JMIR Medical Informatics

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

Contents

Reviews

Viewpoints

Implementation Reports

Original Papers

Corrigenda and Addendas

Machine Learning Models for Parkinson Disease: Systematic Review

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

1

2

3

4

Corresponding Author:
Thasina Tabashum, MSc

Abstract

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

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

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

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

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

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

KEYWORDS

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

Introduction

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

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

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

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

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

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

Methods

Search Strategy

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

Inclusion and Exclusion

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

Questionnaire Design

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

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

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

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

Data Extraction

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

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

Results

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

General Overview of Studies

Methods Applied

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

Data Modalities and Sources

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

ML Modeling Practices

PD Prediction Target

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

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

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

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

Class Imbalance

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

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

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

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

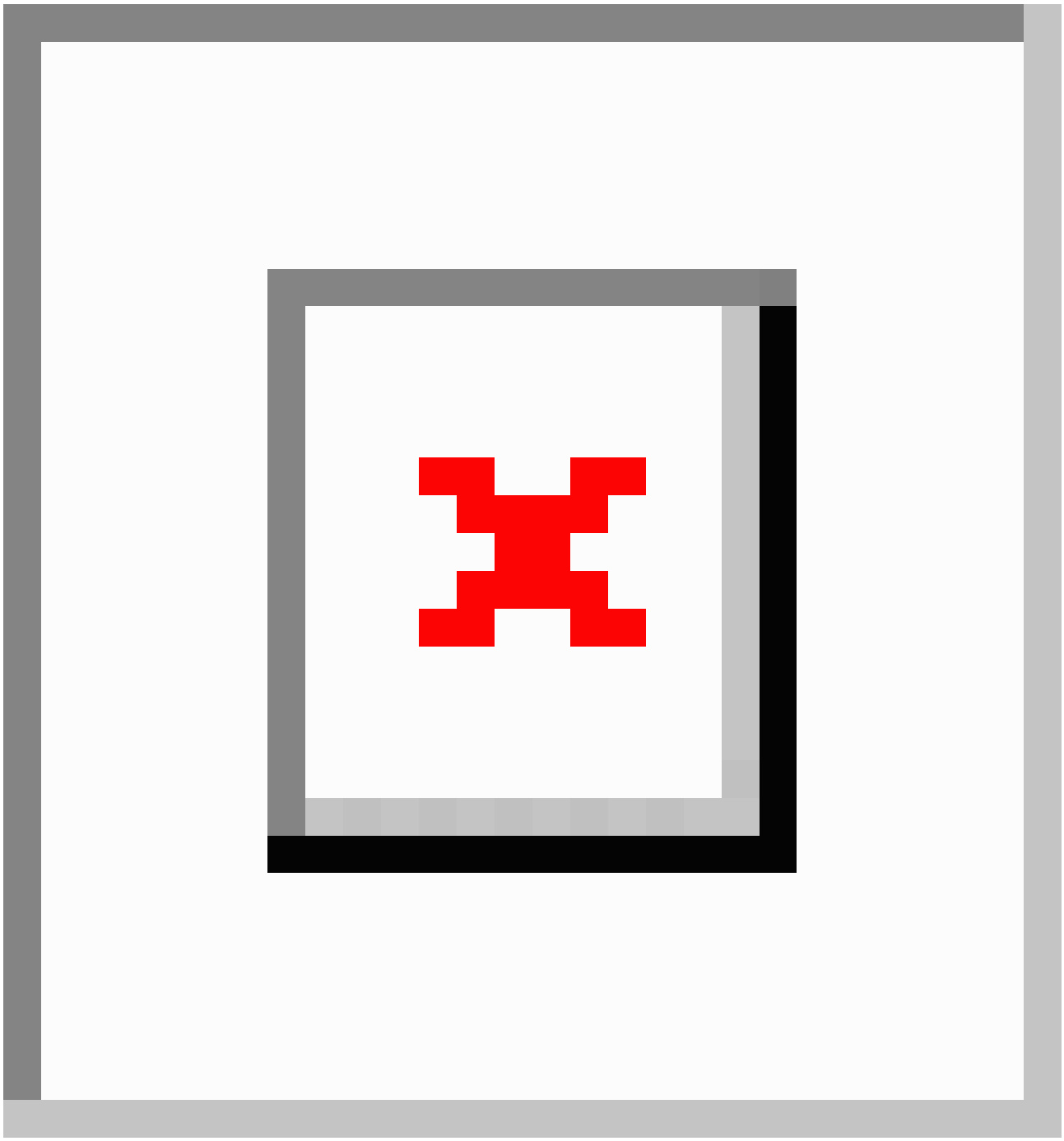


Table . Comparison between imbalanced data versus imbalance mitigation strategies.

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

[a]PD: Parkinson disease.

[b]SMOTE: synthetic minority oversampling technique.

[c]H&Y: Hoehn and Yahr.

Data Set Splitting

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

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

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

Cross-Validation

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

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

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

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

Overfitting

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

Hyperparameters

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

Table . Distribution of studies according to hyperparameter tuning methods.

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

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

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

Model Comparison

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

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

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

Discussion

Principal Findings

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

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

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

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

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

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

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

Conclusion

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

Data Availability

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

Authors' Contributions

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

Conflicts of Interest

None declared.

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

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

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

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

References

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

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

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

Abbreviations

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

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

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

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

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

[1]

[2]

Corresponding Author:
Jiamei Zhou, BA, MA

Abstract

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

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

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

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

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

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

KEYWORDS

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

Introduction

Background

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

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

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

Research Problem and Objective

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

Methods

Overview

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

Search Strategy

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

Table . Search terms used to find studies.

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

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

Eligibility Criteria

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

Study Selection

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

Data Extraction and Synthesis

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

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

Ethical Considerations

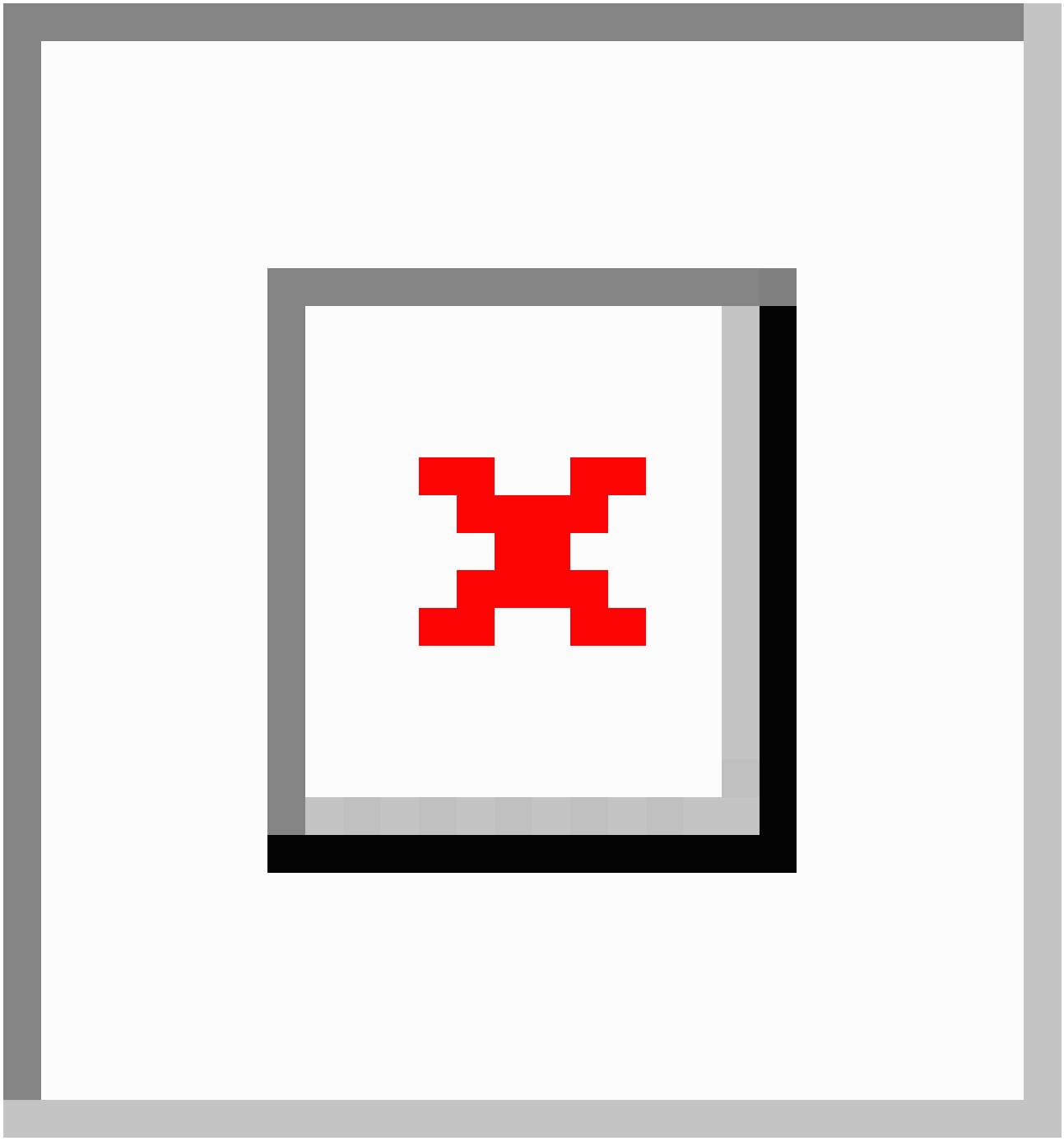
This study did not require ethical approval because we did not study any human or animal subjects and did not collect any personal information or sensitive data.

Results

Search Results

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

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



Characteristics of Included Studies

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

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

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

[a]ICU: intensive care unit.

[b]PICU: pediatric intensive care unit.

[c]NTDB: National Trauma Data Bank.

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

[e]eRI: eICU Research Institute.

AI Technical Characteristics of Included Studies

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

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

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

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

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

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

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

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

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

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

[ak]BUN: blood urea nitrogen.

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

[am]LightGBM: light gradient boosting machine.

[an]COPD: chronic obstructive pulmonary disease.

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

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

Discussion

Principal Findings

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

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

Implications for Practice and Research

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

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

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

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

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

Strengths

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

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

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

Limitations

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

Conclusions

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

Data Availability

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

Authors' Contributions

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

Conflicts of Interest

None declared.

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

References

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

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

Abbreviations

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

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

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

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

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

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

[1]

[2]

[*]all authors contributed equally

Corresponding Author:
Sari Palojoki, PhD

Abstract

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

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

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

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

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

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

KEYWORDS

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

Introduction

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

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

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

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

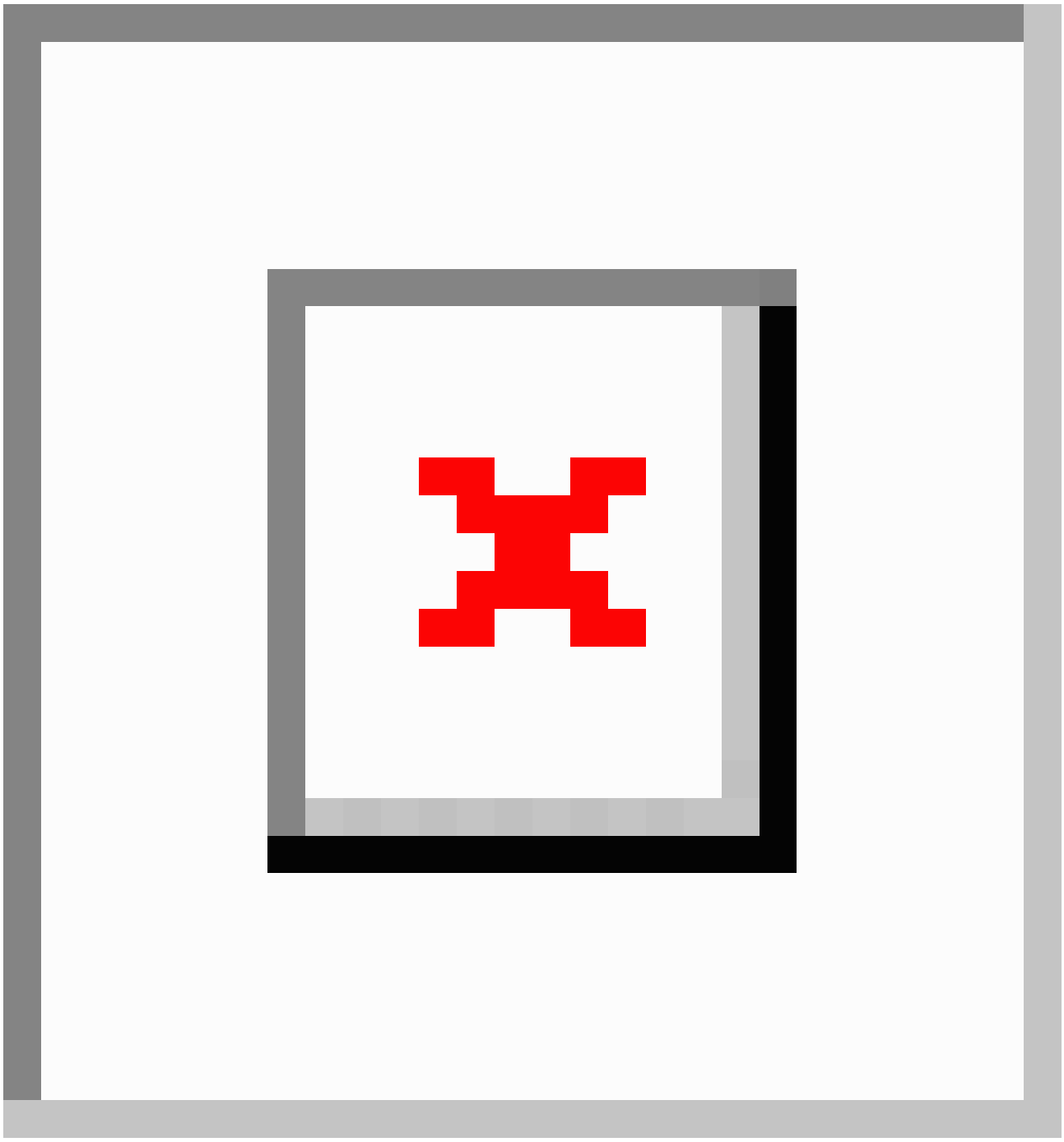
Methods

Methodological Framework

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

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

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



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

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

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

Study Design

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

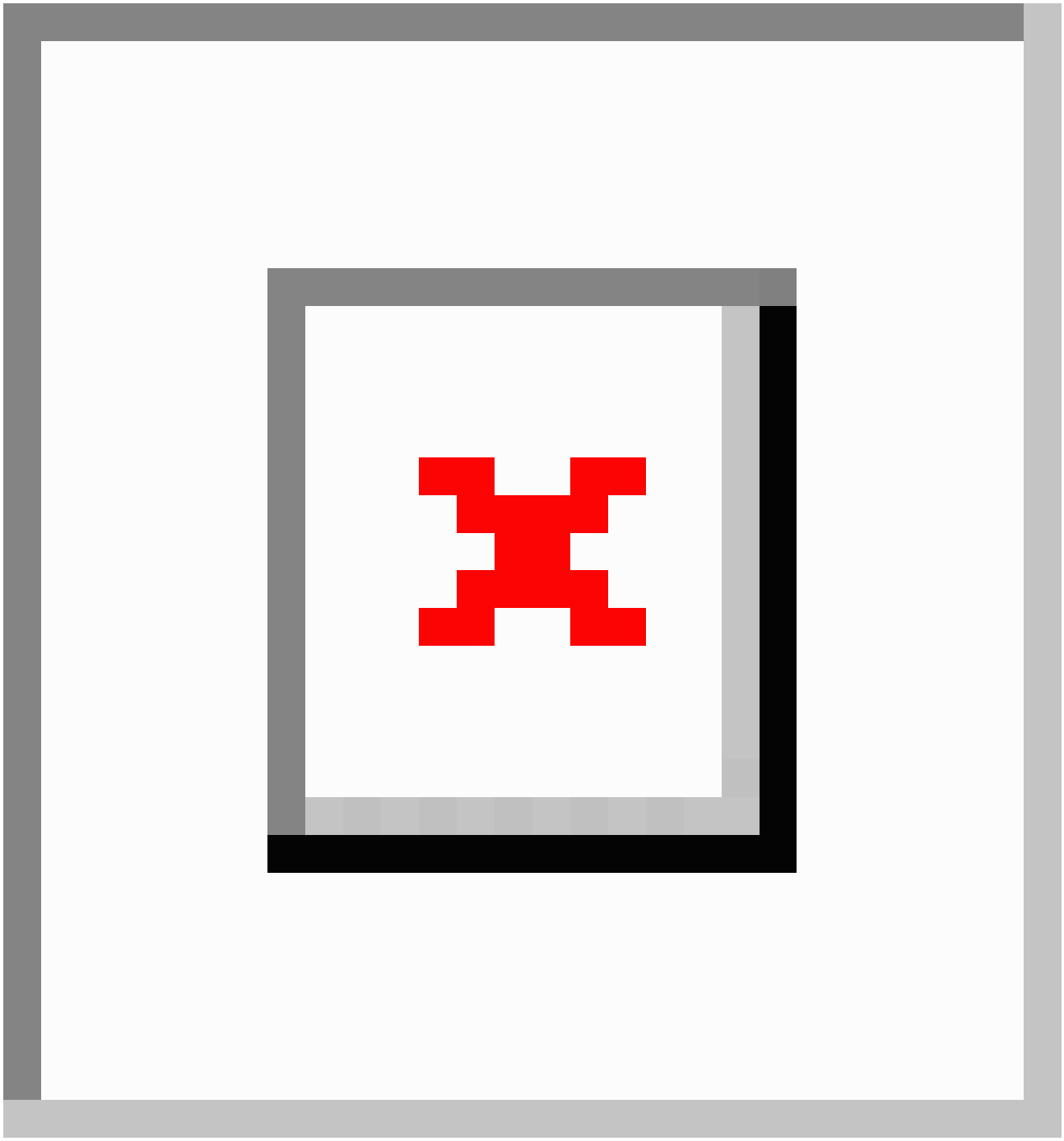
Textbox 1. Search strategy and filters used.

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

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

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

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



Results

Contextual Results

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

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

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

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

Interoperability Results

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

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

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

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

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

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

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

Discussion

Principal Findings

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

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

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

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

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

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

Limitations

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

Suggestions for Future Research

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

Conclusions

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

Acknowledgments

The study was supported by Finnish governmental study grant TYH2021319.

Conflicts of Interest

None declared.

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

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

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

References

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

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

38. 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 Feb 6;11:e43750. [doi: 10.2196/43750] [Medline: 36745498]

Abbreviations

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

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

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

Predicting Hypoxia Using Machine Learning: Systematic Review

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

1

2

3

4

Corresponding Author:
Lena Pigat, MPH

Abstract

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

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

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

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

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

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

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

KEYWORDS

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

Introduction

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

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

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

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

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

Methods

Protocol

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

Search Strategy

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

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

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

Selection Criteria

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

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

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

Data Extraction and Risk of Bias

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

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

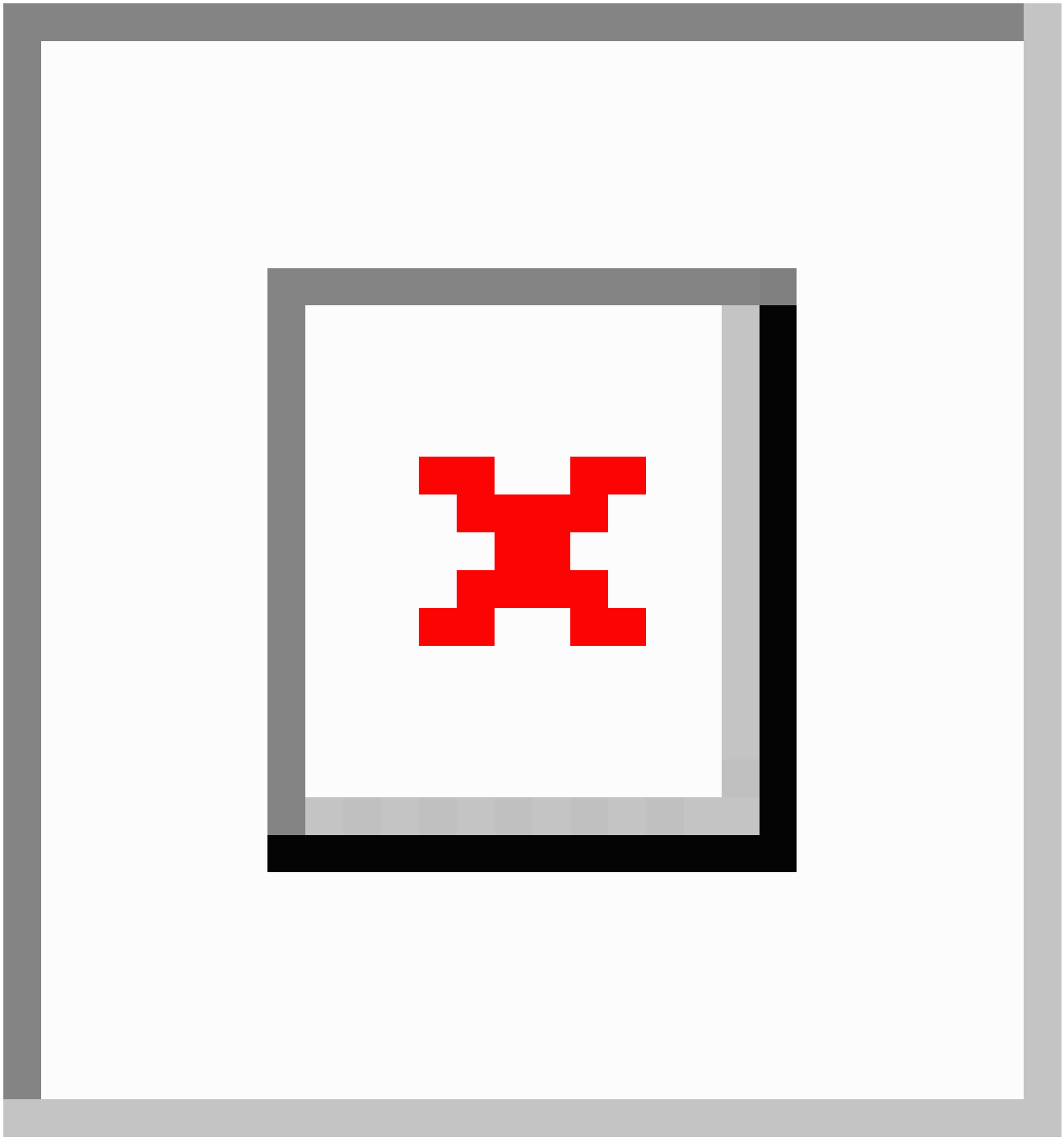
To assess the risk of bias, quality, and applicability of the studies included, Prediction Model Risk of Bias Assessment Tool (PROBAST) was used [14]. This tool is specifically designed to investigate the quality of prediction models and has become increasingly prevalent in systematic reviews in recent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

Results

Literature Search

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

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



Study Characteristics

Overview

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

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

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

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

[a]SpO_2: peripheral oxygen saturation.

[b]LSTM: long short-term memory.

[c]PPV: positive predictive value.

[d]SaO_2: arterial oxygen saturation.

[e]GBT: gradient boosted tree.

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

[g]Lin: linear regression.

[h]LR: logistic regression.

[i]ANN: artificial neural network.

[j]XGB: extreme gradient boosting.

[k]RNN: recurrent neural network.

[l]GBM: gradient boosting machine.

[m]PaO_2: partial pressure of oxygen.

[n]FiO_2: fraction of inspired oxygen.

[o]NN: neural network.

[p]RF: random forest.

[q]HR: heart rate.

[r]BR: breath rate.

[s]SBP: systolic blood pressure.

[t]DBP: diastolic blood pressure.

[u]AG: anion gap.

[v]CNN: convolutional neural network.

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

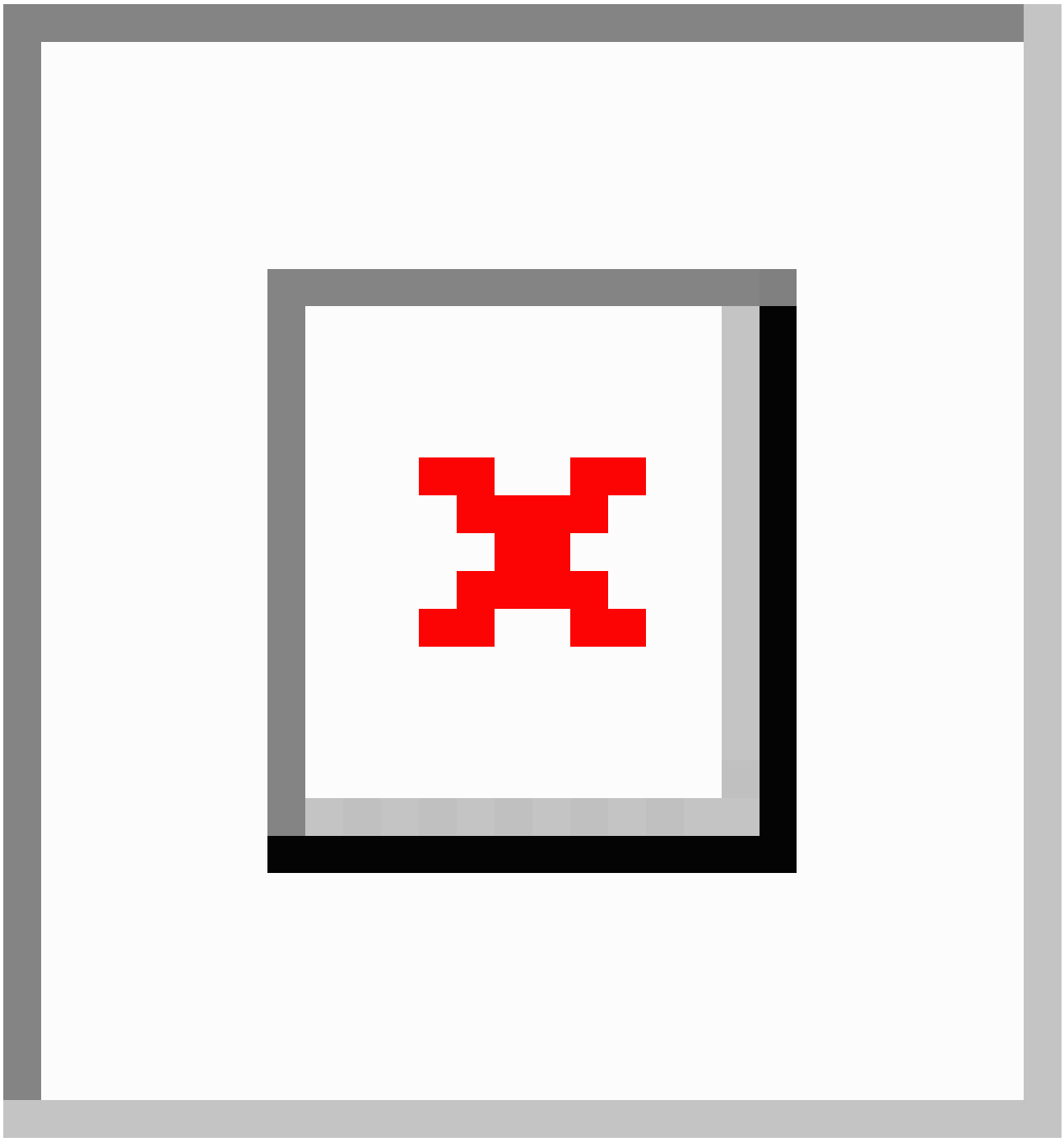
Data Sources and Population

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

ML Model Specifics

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

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



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

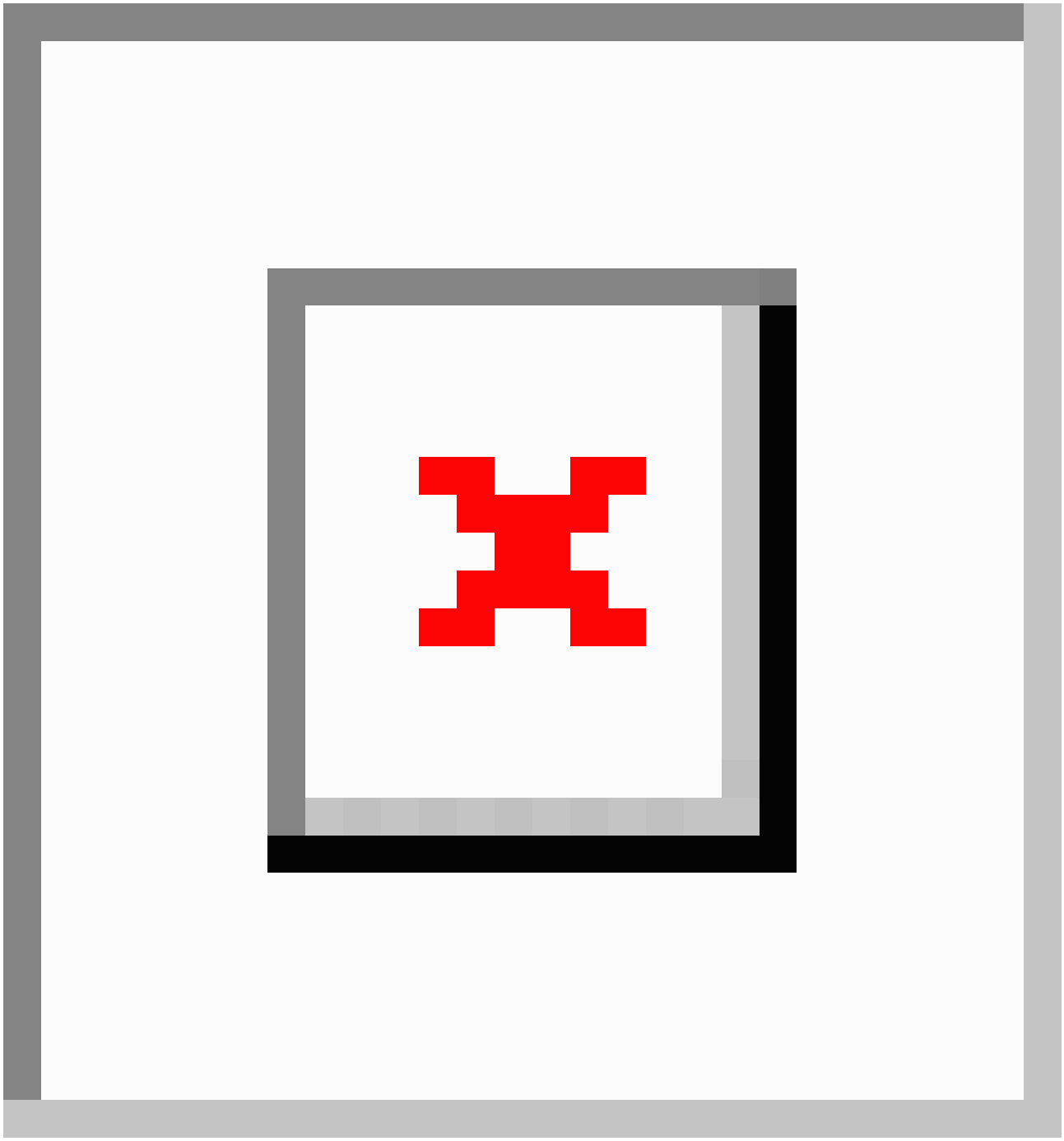
The prediction end point was defined by a threshold of SpO_2 between 89% and 92% for most of the studies (7/12, 58%) [15,17-20,23,24]. Thresholds of the partial pressure of oxygen, the arterial oxygen saturation, or the ratio of partial pressure of oxygen to the fraction of inspired oxygen were used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

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

Performance

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

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



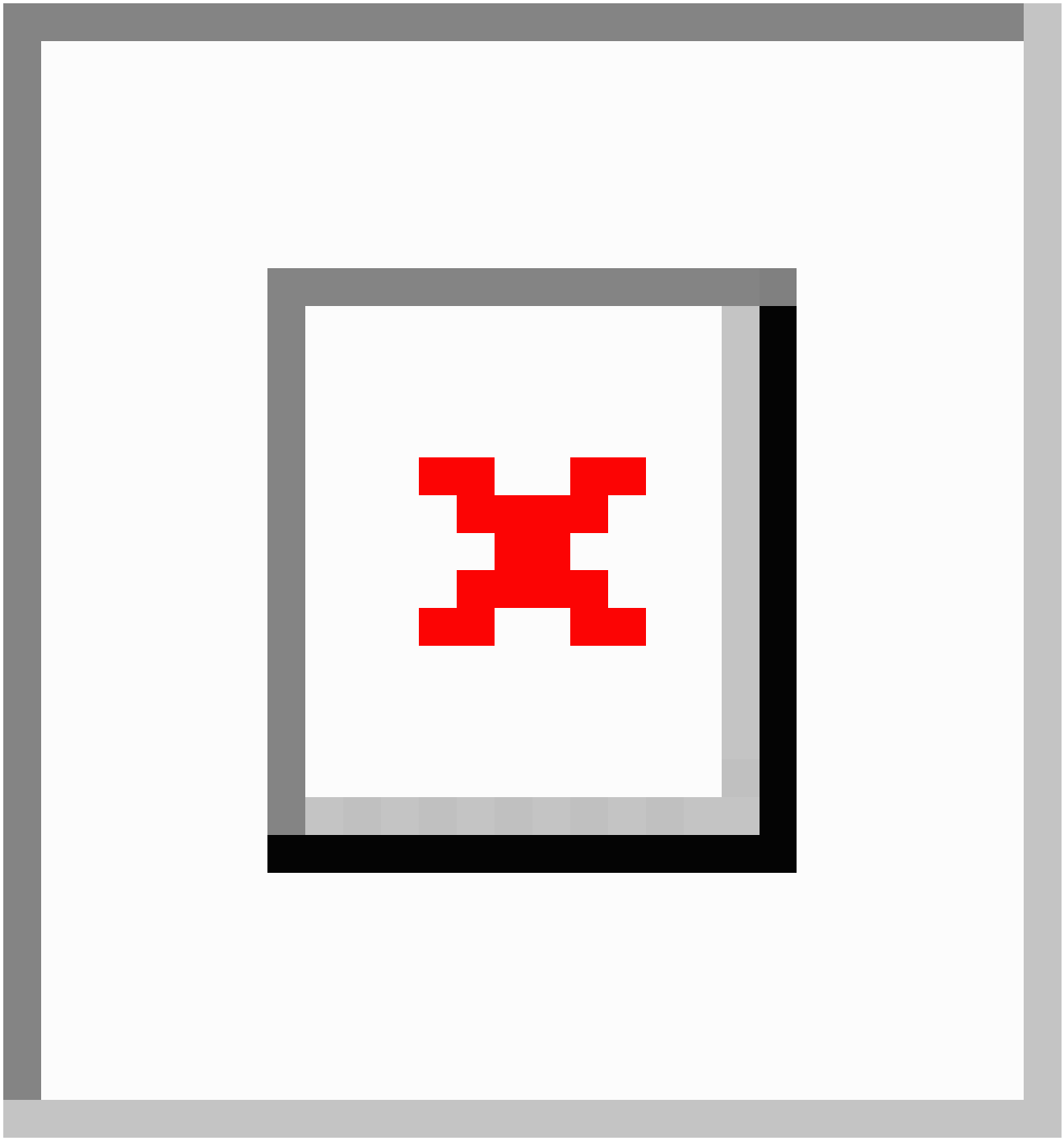
Of the 9 studies reporting AUROC, 8 (89%) showed a value higher than 0.75 [16-18,20-24]. This included 3 studies that showed a significant trade-off between sensitivity and specificity [17,21,24]. The overall performance was moderate or high with respect to classification metrics, both in studies performing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

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

Risk-of-Bias Assessment

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

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



Discussion

Principal Findings

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

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

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

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

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

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

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

Applicability and Future Opportunities

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

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

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

Limitations

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

Conclusion

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

Acknowledgments

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

Authors' Contributions

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

Conflicts of Interest

None declared.

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

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

References

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

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

Abbreviations

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

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

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

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

Review

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

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

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

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

[*]all authors contributed equally

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

Abstract

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

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

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

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

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

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

KEYWORDS

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

Introduction

Overview

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

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

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

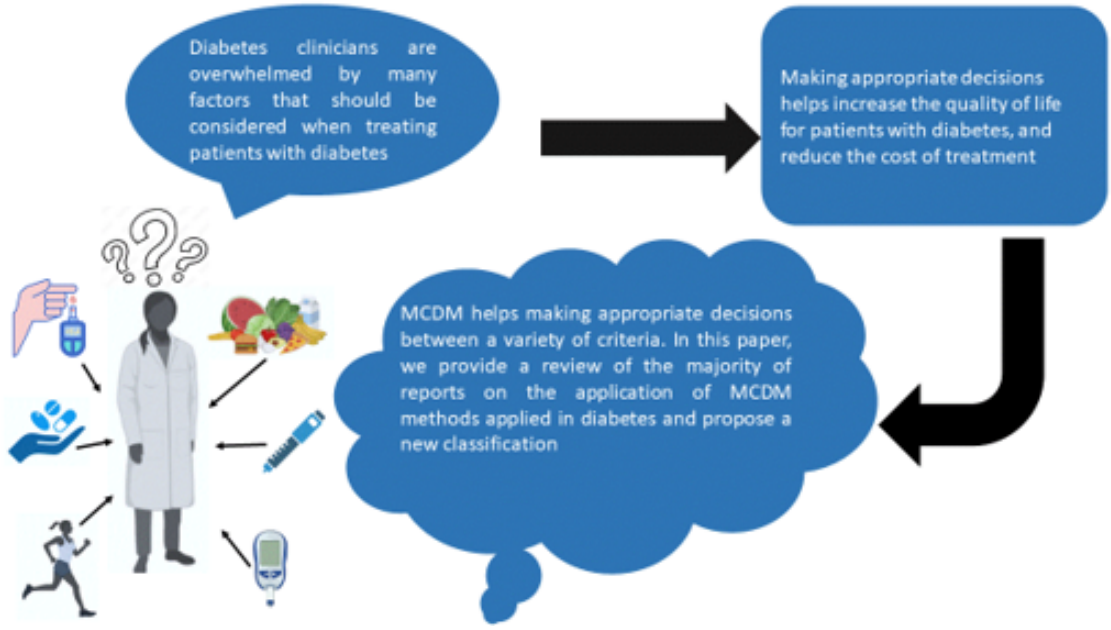
In addition to making informed decisions about the patient's health, endocrinologists and general practitioners should carefully assess various factors, including lifestyle choices, dietary habits, daily physical activity levels, insulin requirements, and the patient's willingness to embrace self-management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

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

Objective

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

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



Methods

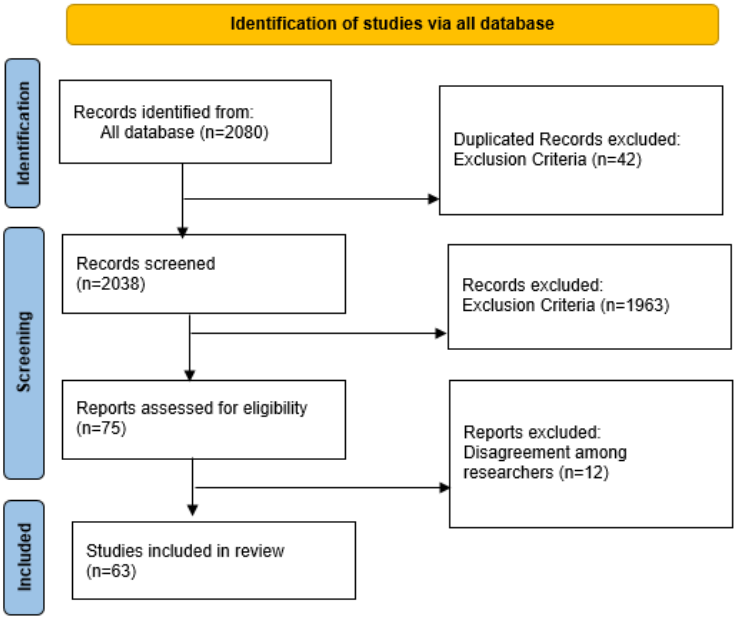
Search Strategies

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

Inclusion and Exclusion Criteria

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

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



Results

Overview

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

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

MCDM Techniques Overview

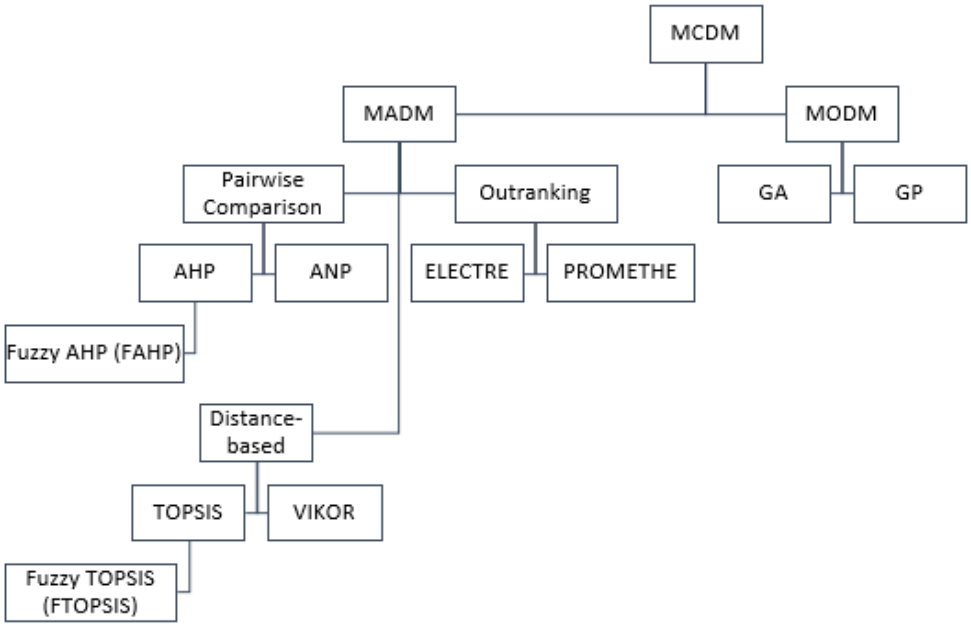
Since so many choices in our modern lives depend on a multitude of factors, the decision can be made by giving various criteria varying weights, which is done by expert groups. Determining the structure and explicitly evaluating several criteria is crucial. Therefore, constructing and resolving multicriteria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

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

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

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

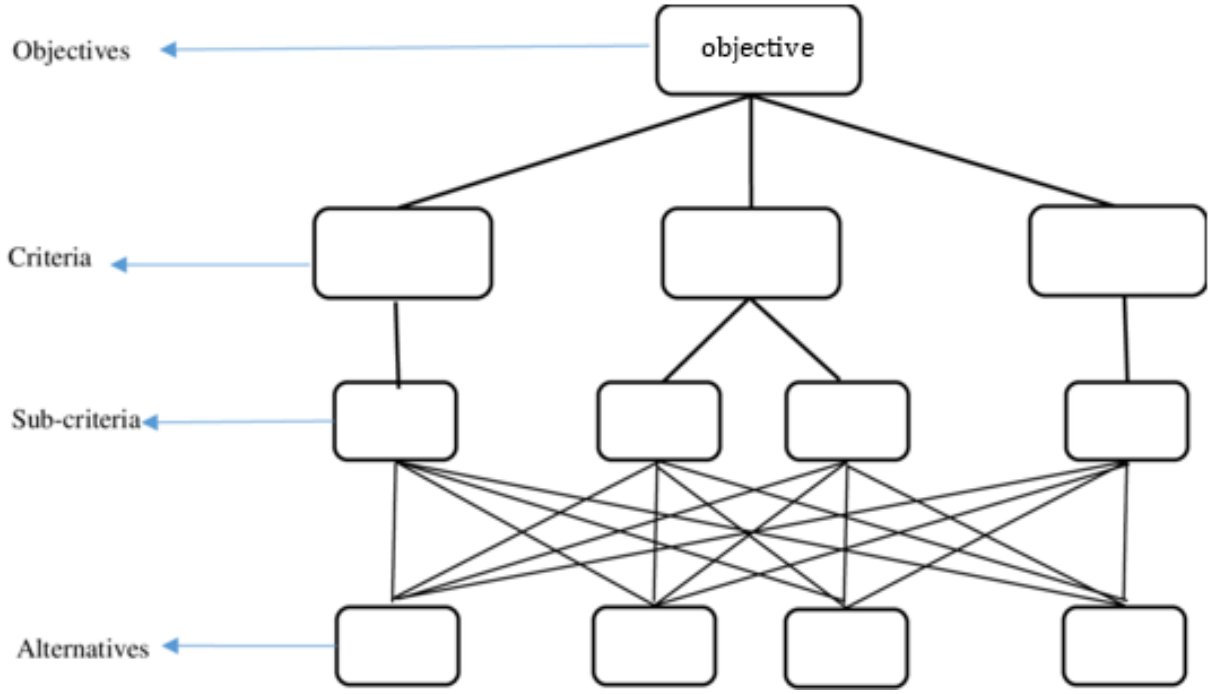
Figure 3. Hierarchical structures of MCDM methods. AHP: analytic hierarchy process; ANP: analytical network process; ELECTRE: Elimination Et Choix Traduisant la Realité; GA: genetic algorithm; GP: goal programming; MADM: multiattribute decision-making; MCDM: multicriteria decision-making; MODM: multiobjective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



AHP Method

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

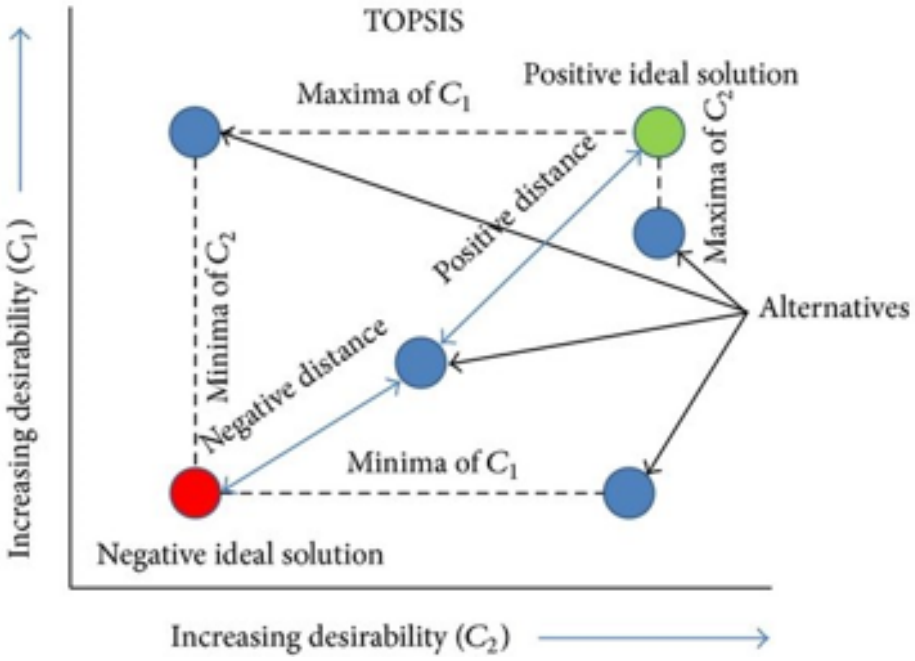
Figure 4. Hierarchical structure of analytic hierarchy process.



TOPSIS Method

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

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



ANP Method

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

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

Weighting Methods

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

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

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

Selection of Diabetes Medication

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

Table 1. Diabetes medication publications.

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

[a]AHP: analytic hierarchy process.

[b]T2D: type 2 diabetes.

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

[d]FMULTIMOORA: full multiplicative form.

[e]WASPAS: Weighted Aggregated Sum Product Assessment.

[f]CRITIC: Criteria Importance Through Intercriteria Correlation.

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

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

Diagnosis of Diabetes

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

Table 2. Diabetes diagnosis publications.

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

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

[b]BP: blood pressure.

[c]ACR: albumin creatinine ratio.

[d]AHP: analytic hierarchy process.

[e]FAHP: fuzzy analytic hierarchy process.

[f]GDM: gestational diabetes mellitus.

[g]N/A: not applicable.

[h]T2D: type 2 diabetes.

[i]FBS: fasting blood sugar.

[j]PRF: physical risk factors.

[k]DIBAR: Created Diabetes Risk Assessment.

[l]FRBS: fuzzy rule-based systems.

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

[n]MCDM: multicriteria decision-making.

[o]WPM: Weighted Product Model.

Meal Recommendation for Diabetes

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

Table 3. Meal recommendation publications.

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

[a]AHP: analytic hierarchy process.

[b]N/A: not applicable.

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

[d]ANP: analytical network process.

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

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

Diabetes Management

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

Table 4. Diabetes management publications.

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

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

[b]VIKOR: ViseKriterijumska Optimizacija I Kompromisno Resenje.

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

[d]mHealth: mobile health.

[e]T2D: type 2 diabetes.

[f]ANP: analytical network process.

[g]CRITIC: Criteria Importance Through Intercriteria Correlation

[h]T2DM: type 2 diabetes mellitus.

[i]AHP: analytic hierarchy process.

[j]SCP: Supply Chain Partnership.

[k]SLP: Systematic Layout Planning.

[l]HbA1c: hemoglobin A1c.

[m]LDL: low-density lipoprotein.

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

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

Diabetes Complication

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

Table 5. Diabetes complication diagnosis publications.

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

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

[b]BP: blood pressure.

[c]MCDM: multicriteria decision-making.

[d]GFR: estimated glomerular filtration rate.

[e]CKD: chronic kidney disease.

[f]DR: diabetic retinopathy.

[g]AUC: area under the curve.

[h]TPR: true positive rate.

[i]FNR: false negative rate.

[j]FPR: false positive rate.

[k]TNR: true negative rate.

[l]VIKOR: ViseKriterijumska Optimizacija I Kompromisno Resenje.

[m]HDL: high-density lipoprotein.

[n]AHP: analytic hierarchy process.

[o]DBP: diastolic blood pressure.

[p]N/A: not applicable.

Discussion

Principal Findings

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

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

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

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

Conclusions

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

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

Acknowledgments

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

Conflicts of Interest

None declared.

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

References

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

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

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

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

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

Abbreviations

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

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

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

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

Review

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

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

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

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

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

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

Abstract

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

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

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

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

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

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

KEYWORDS

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

Introduction

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

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

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

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

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

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

Methods

Overview

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

Textbox 1. Search query used.

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

Inclusion and Exclusion Criteria

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

Selection of Articles

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

Data Extraction

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

Data Synthesis

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

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

Data quality dimension and definition

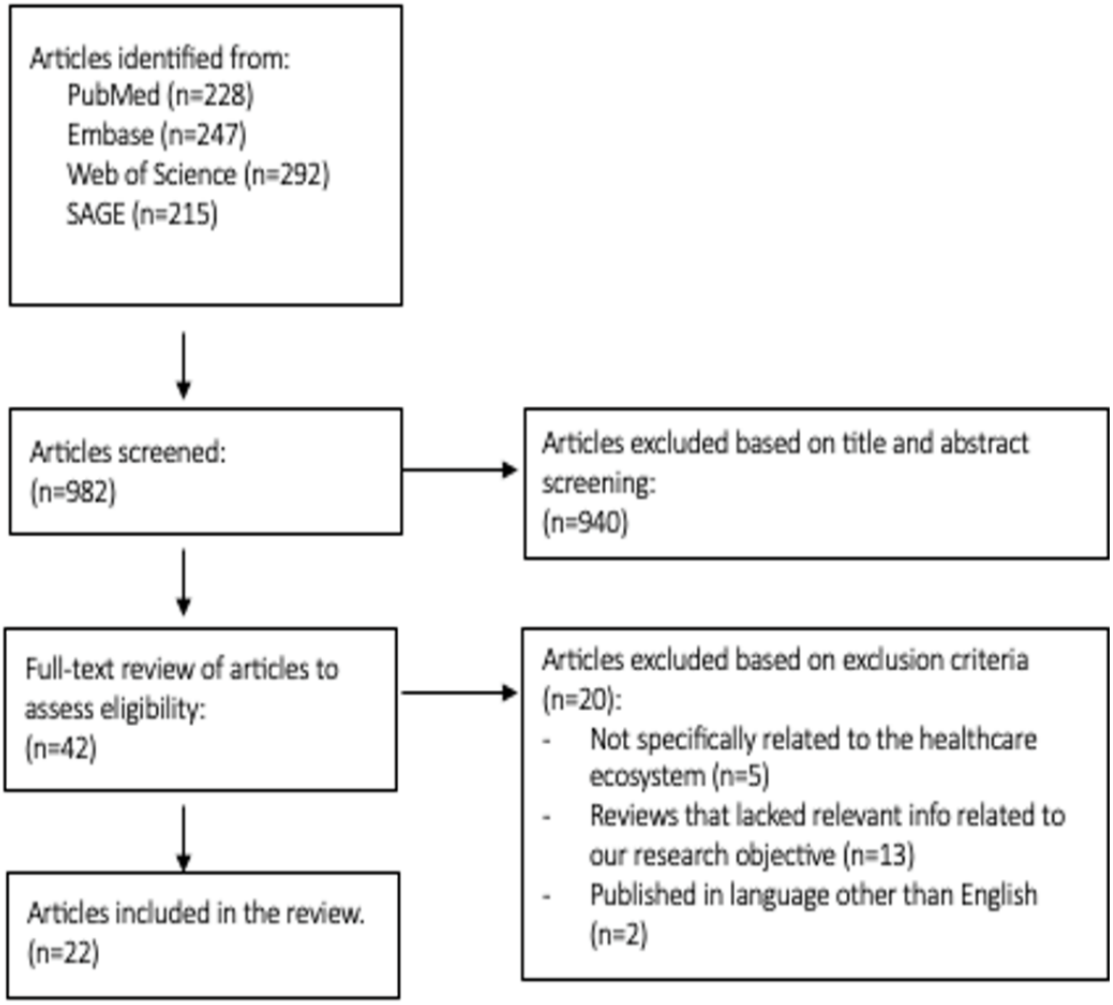
- Completeness: the extent to which data are present
- Consistency: the extent to which data satisfy constraints
- Correctness: the extent to which data are true and unbiased
- Timeliness: the extent to which data are promptly processed and up to date
- Stability: the extent to which data are comparable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

Results

Search Process

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

Figure 1. The process of selecting articles.



Data Sources

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

Observed Frameworks for Data Quality Dimensions

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

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

Overview

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

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

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

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

[a]EHR: electronic health record.

Completeness

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

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

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

Consistency

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

Correctness

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

Timeliness

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

Stability

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

Contextualization

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

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

Representation

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

Trustworthiness

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

Uniqueness

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

Observed Data Quality Assessment Methods

Overview

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

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

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

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

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

[a]M: method.

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

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

Completeness

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

Consistency

- Conformance [24]

- M18: element presence
- M17: conformance check

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

Correctness

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

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

Stability

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

Timeliness

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

Trustworthiness

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

Representation

- Relevance [39]
  - M38: not specified

Completeness

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

Consistency

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

Correctness

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

Stability

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

Timeliness

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

Trustworthiness

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

Representation

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

Uniqueness and Contextualization

No assessment methods were mentioned for these data quality dimensions.

Discussion

Principal Findings

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

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

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

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

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

Limitations

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

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

Conclusions

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

Conflicts of Interest

None declared.

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

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

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

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

References

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

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

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

Abbreviations

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

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

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

Review

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

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

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

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

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

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

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

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

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

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

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

Abstract

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

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

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

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

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

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

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

KEYWORDS

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

Introduction

Diabetes and Nutrition

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

Telehealth and Diabetes Self-Management

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

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

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

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

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

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

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

Methods

Overview

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

Identifying the Research Questions

This review seeks to answer the following research questions:

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

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

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

Identifying and Selecting the Studies

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

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

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

Charting the Data

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

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

Collecting, Summarizing, and Reporting the Results

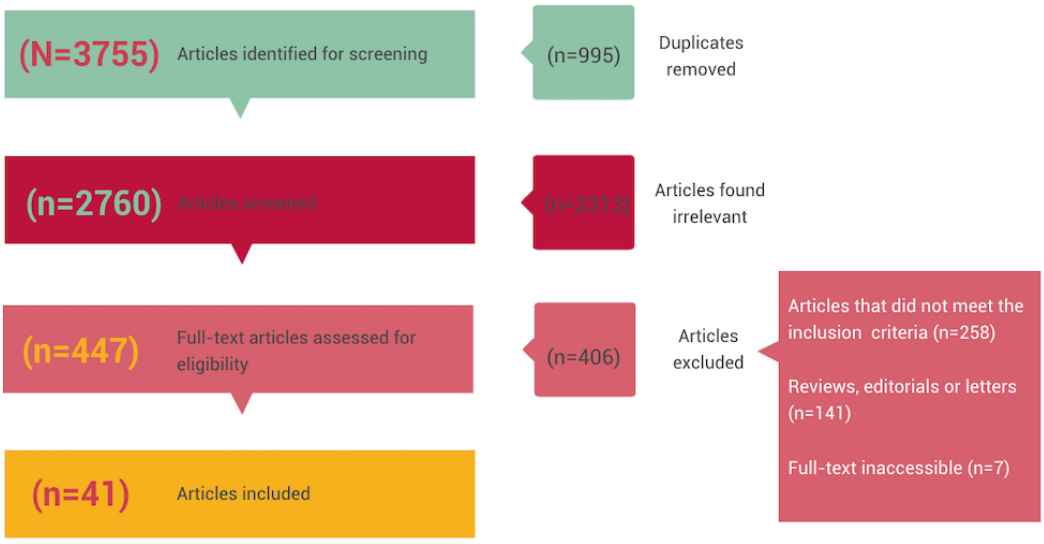
Again according to the framework developed by Arksey and O'Malley [24] and the revisions by Levac et al [25], descriptive web-based abstracts and thematic analyses performed with NVivo software (release 1.7; Lumivero) were used for data analysis, yielding an approach resembling that of a narrative review.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

Results

Overview

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

Figure 1. Flow diagram of study selection.



Characteristics of the Studies

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

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

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

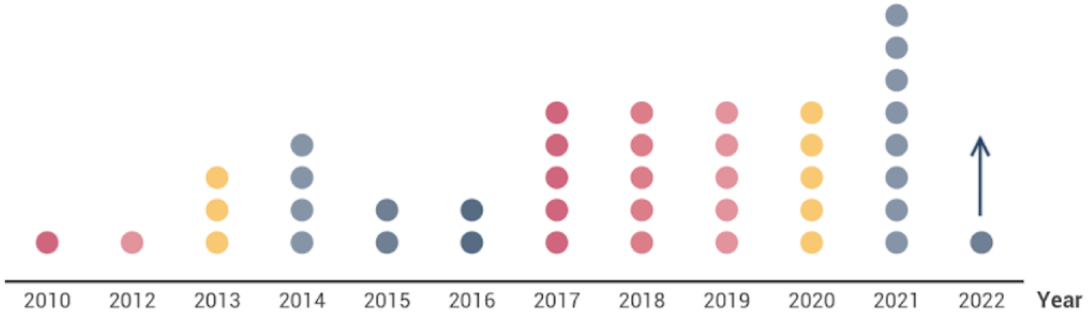
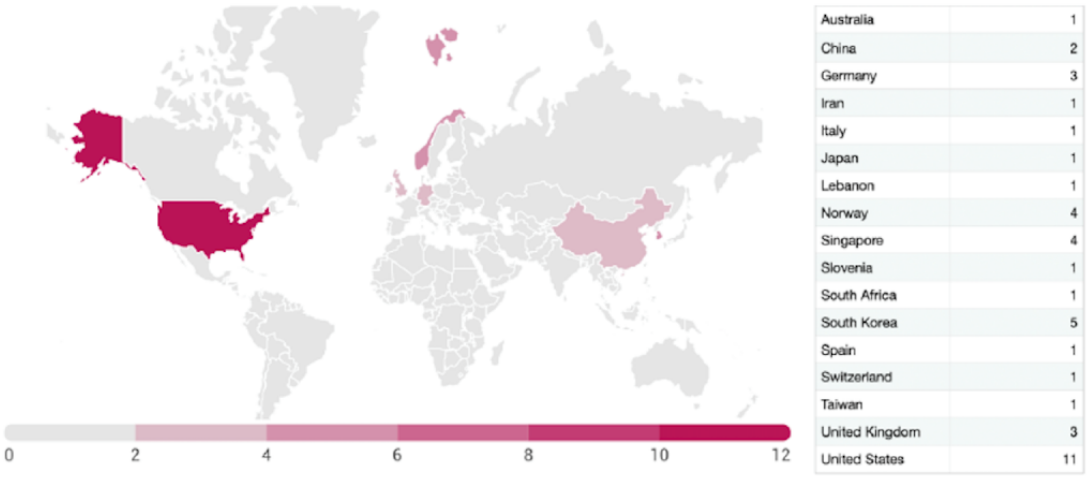


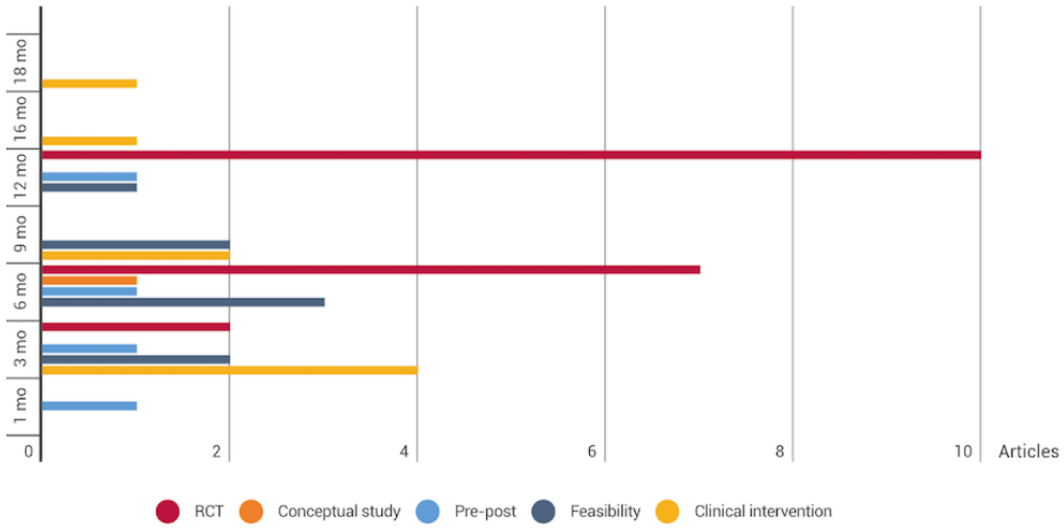
Figure 3. Countries in which the studies were published. The dots represent articles and the x-axis denotes the years.



General Characteristics of the Intervention

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

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



Health Professional Involvement

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

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

Characteristics of Digital Self-Management

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

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

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

Key Digital Tool Features With Positive Impacts on the Health Condition

Monitoring of Glycemia Levels

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

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

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

Diet Monitoring

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

Physical Activity Monitoring

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

Medication Monitoring

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

Therapeutic Education

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

Health Professional Involvement

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

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

Other Health Data Monitoring

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

Health Care Management

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

Combination of Interventions

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

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

Discussion

Principal Findings

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

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

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

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

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

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

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

Recommendations for Future Designs

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

Limitations of Included Studies

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

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

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

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

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

Strengths and Limitations of This Review

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

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

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

Future Research Prospects

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

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

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

Conclusions

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

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), 517 KB - medinform_v12i1e46699_app1.pdf]

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

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

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

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

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

References

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

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

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

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

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

Abbreviations

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

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

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

Review

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

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

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

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

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

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

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

Abstract

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

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

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

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

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

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

KEYWORDS

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

Introduction

Background

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

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

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

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

Motivation

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

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

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

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

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

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

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

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

[a]Definitions.

[b]Computational methods and approaches.

[c]Evaluation metrics.

Methods

Research Methodology

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

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

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

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

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

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

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

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

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

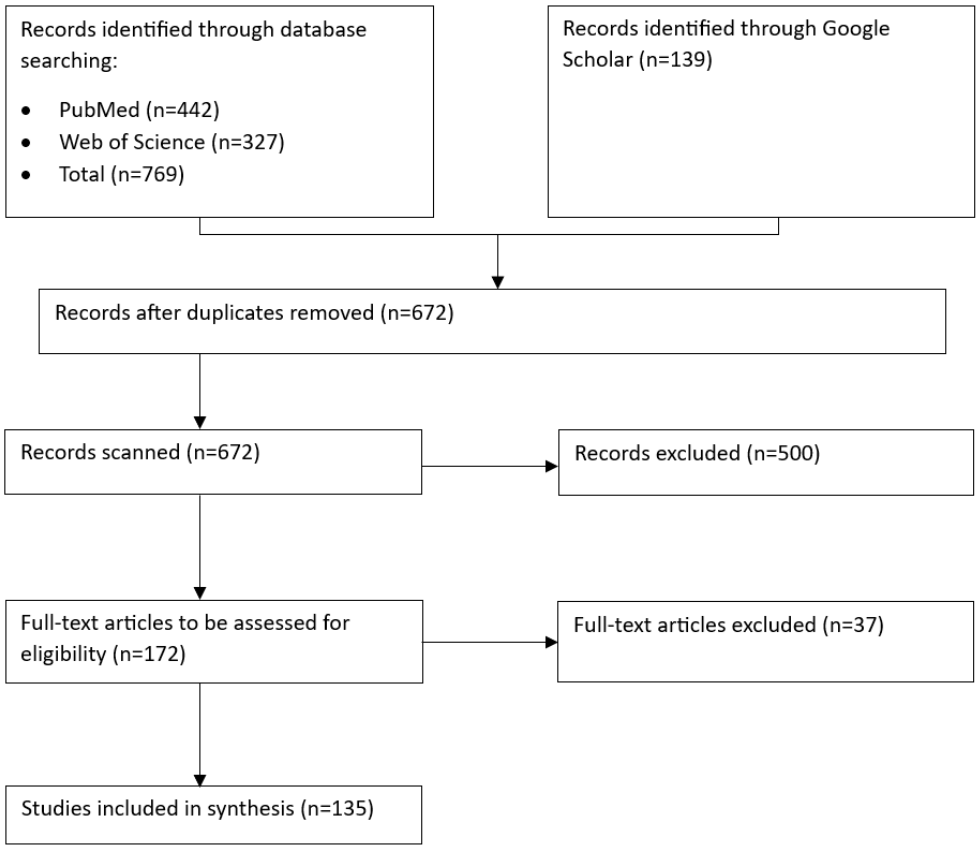


Table 2. Search strategy for finding research articles.

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

[a]G1: group 1.

[b]G2: group 2.

[c]G3: group 3.

[d]T1: search term 1.

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

[f]T2: search term 2.

[g]N/A: not applicable.

Textbox 1. The initial query to the databases.

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

Textbox 2. Modified query to the databases.

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

Data Items and Data-Charting

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

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

Results

Definitions, Computational Methods, and Approaches to Fairness

Overview

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

Calibrated Fairness

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

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

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

Statistical Fairness

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

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

Intersectional Fairness

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

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

Definitions, Computational Methods, and Approaches to Accountability

Overview

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

Legal Accountability

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

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

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

Ethical Accountability

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

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

Technical Accountability

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

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

Societal Accountability

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

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

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

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

Definitions, Computational Methods, and Approaches to Transparency

Overview

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

Algorithmic Transparency

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

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

Data Transparency

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

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

Process Transparency

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

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

Explainability and Interpretability

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

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

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

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

Definitions, Computational Methods, and Approaches to Ethics

Overview

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

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

Philosophical Ethics

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

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

Professional Ethics

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

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

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

Legal Ethics

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

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

Other Ethical Considerations

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

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

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

Discussion

Principal Findings and Future Research Directions

Overview

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

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

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

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

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

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

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

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

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

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

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

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

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

Limitations

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

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

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

Conclusions

Our review sheds light on the current state of FATE in health care AI as applied to SMPs. It offers a critical analysis of the methodologies, computational techniques, and evaluative strategies used in various studies. By examining the successes and identifying the shortcomings of current practices, this review stimulates further innovation in the field. It challenges existing paradigms on how AI technologies can be both technologically advanced and ethically robust, ensuring fairness, accountability, and transparency in their application.

The practical implications of this work are substantial. First, it guides future research by identifying recent trends and research gaps, suggesting that researchers focus on creating more robust, fair, and ethical AI systems. This includes using diverse data sets that more accurately represent the global population and using evaluation metrics that comprehensively assess the systems' impacts on all stakeholders. Second, this review underscores the importance of integrating FATE principles throughout the AI system development life cycle, from conceptualization to deployment. For practitioners in health care and technology, this signifies a move toward more inclusive, transparent, and ethically guided development processes. Such a transition not only addresses biases and accountability issues but also boosts patient trust and engagement with AI-driven health care solutions on social media.

Third, the insights from this review are invaluable for policy makers and regulatory bodies, aiding in the creation of nuanced regulations and guidelines that ensure that AI technologies positively contribute to health care outcomes without compromising ethical standards or patient rights. Furthermore, by simplifying complex concepts, this review acts as an educational tool for a broad audience, including health care providers, AI developers, patients, and the general public. Raising awareness about the importance of FATE in health care AI fosters more informed participation in discussions and decision-making regarding AI use in health care, particularly on SMPs.

Ultimately, this study aids in the pursuit of ethical development and deployment of AI systems in health care. By providing an in-depth analysis of the current achievements and future directions for FATE in health care AI on social media, it advocates for the adoption of best practices that balance ethical considerations with technological innovations. The implications of this study extend beyond academia, affecting how AI technologies are conceptualized, developed, and implemented in health care on social media, thereby shaping a future where AI-driven health care solutions are not only effective and innovative but also ethically responsible, equitable, and transparent. This ensures that these technologies serve the best interests of society.

Data Availability

All data generated or analyzed during this study are included in this published paper and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Studies selected for the review.
[DOCX File , 29 KB - medinform_v12i1e50048_app1.docx]

Multimedia Appendix 2
Fairness evaluation metrics with mathematical formulation.
[DOCX File , 27 KB - medinform_v12i1e50048_app2.docx]

Multimedia Appendix 3
Accountability evaluation metrics with mathematical formulation.

[DOCX File , 20 KB - medinform_v12i1e50048_app3.docx]

Multimedia Appendix 4
Transparency evaluation metrics with mathematical formulation.
[DOCX File , 20 KB - medinform_v12i1e50048_app4.docx]

Multimedia Appendix 5
Ethics evaluation metrics with mathematical formulation.
[DOCX File , 18 KB - medinform_v12i1e50048_app5.docx]

References

1. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021 Jul 13;54(6):1-35. [doi: 10.1145/3457607]
2. Mashhadi A, Winder SG, Lia EH, Wood SA. No walk in the park: the viability and fairness of social media analysis for parks and recreational policy making. Proc Int AAAI Conf Web Soc Media 2021 May 22;15(1):409-420. [doi: 10.1609/icwsm.v15i1.18071]
3. Leonelli S, Lovell R, Wheeler BW, Fleming L, Williams H. From FAIR data to fair data use: methodological data fairness in health-related social media research. Big Data Soc 2021 May 03;8(1). [doi: 10.1177/20539517211010310]
4. Singhal A, Baxi MK, Mago V. Synergy between public and private health care organizations during COVID-19 on Twitter: sentiment and engagement analysis using forecasting models. JMIR Med Inform 2022 Aug 18;10(8):e37829 [FREE Full text] [doi: 10.2196/37829] [Medline: 35849795]
5. Kington RS, Arnesen S, Chou WY, Curry SJ, Lazer D, Villarruel AM. Identifying credible sources of health information in social media: principles and attributes. NAM Perspect 2021;2021:10.31478/202107a [FREE Full text] [doi: 10.31478/202107a] [Medline: 34611600]
6. Pershad Y, Hangge PT, Albadawi H, Oklu R. Social medicine: Twitter in healthcare. J Clin Med 2018 May 28;7(6):121 [FREE Full text] [doi: 10.3390/jcm7060121] [Medline: 29843360]
7. Flores L, Young SD. Ethical considerations in the application of artificial intelligence to monitor social media for COVID-19 data. Minds Mach (Dordr) 2022;32(4):759-768 [FREE Full text] [doi: 10.1007/s11023-022-09610-0] [Medline: 36042870]
8. Pirraglia PA, Kravitz RL. Social media: new opportunities, new ethical concerns. J Gen Intern Med 2013 Feb 8;28(2):165-166 [FREE Full text] [doi: 10.1007/s11606-012-2288-x] [Medline: 23225258]
9. Wieringa M. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020 Presented at: FAT* '20; January 27-30, 2020; Barcelona, Spain URL: https://dl.acm.org/doi/abs/10.1145/3351095.3372833 [doi: 10.1145/3351095.3372833]
10. Hutchinson B, Smart A, Hanna A, Denton E, Greer C, Kjartansson O, et al. Towards accountability for machine learning datasets: practices from software engineering and infrastructure. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual event, Canada. [doi: 10.1145/3442188.3445918]
11. Johnson SL. AI, machine learning, and ethics in health care. J Leg Med 2019;39(4):427-441. [doi: 10.1080/01947648.2019.1690604] [Medline: 31940250]
12. 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]
13. The Belmont report. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html [accessed 2023-12-05]
14. Shin D. The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. Int J Hum Comput Stud 2021 Feb;146:102551. [doi: 10.1016/j.ijhcs.2020.102551]
15. Hagerty A, Rubinov I. Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv Preprint posted online July 18, 2019 [FREE Full text] [doi: 10.48550/arXiv.1907.07892]
16. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018 Presented at: DSAA 2018; October 1-3, 2018; Turin, Italy. [doi: 10.1109/dsaa.2018.00018]
17. Chakraborty S, Tomsett R, Raghavendra R, Harborne D, Alzantot M, Cerutti F, et al. Interpretability of deep learning models: a survey of results. In: Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation. 2017 Presented at: SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017; August 4-8, 2017; San Francisco, CA. [doi: 10.1109/uic-atc.2017.8397411]

18. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. Electronics 2019 Jul 26;8(8):832. [doi: 10.3390/electronics8080832]
19. Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res 2017 Jun 06;19(6):e195 [FREE Full text] [doi: 10.2196/jmir.7082] [Medline: 28588006]
20. Bear Don't Walk OJ4, Reyes Nieva H, Lee SS, Elhadad N. A scoping review of ethics considerations in clinical natural language processing. JAMIA Open 2022 Jul;5(2):ooac039 [FREE Full text] [doi: 10.1093/jamiaopen/ooac039] [Medline: 35663112]
21. Attard-Frost B, De los Ríos A, Walters DR. The ethics of AI business practices: a review of 47 AI ethics guidelines. AI Ethics 2022 Apr 13;3(2):389-406. [doi: 10.1007/s43681-022-00156-6]
22. 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]
23. Vian T, Kohler JC. Medicines Transparency Alliance (MeTA): pathways to transparency, accountability, and access: cross-case analysis and review of phase II. World Health Organization. 2016 May 25. URL: https://tinyurl.com/3vhjyysd [accessed 2023-12-05]
24. Kofod-Petersen A. How to do a structured literature review in computer science. Norwegian University of Science and Technology. 2018. URL: https://research.idi.ntnu.no/aimasters/files/SLR_HowTo2018.pdf [accessed 2024-03-13]
25. 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]
26. Saha D, Schumann C, Mcelfresh DC, Dickerson JP, Mazurek ML, Tschantz MC. Measuring non-expert comprehension of machine learning fairness metrics. In: Proceedings of the 37th International Conference on Machine Learning. 2020 Presented at: PMLR 2020; July 13-18, 2020; Virtual event URL: https://proceedings.mlr.press/v119/saha20c.html [doi: 10.1145/3375627.3375819]
27. Mehrabi N, Gupta U, Morstatter F, Steeg GV, Galstyan A. Attributing fair decisions with attention interventions. In: Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022). 2022 Presented at: TrustNLP 2022; July 14, 2022; Seattle, WA. [doi: 10.18653/v1/2022.trustnlp-1.2]
28. Hertweck C, Heitz C, Loi M. On the moral justification of statistical parity. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021 Presented at: FAccT '21; March 3-10, 2021; Virtual event. [doi: 10.1145/3442188.3445936]
29. Yao H, Chen Y, Ye Q, Jin X, Ren X. Refining language models with compositional explanations. arXiv Preprint posted online March 18, 2021 [FREE Full text]
30. Markoulidakis I, Kopsiaftis G, Rallis I, Georgoulas I. Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. In: Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference. 2021 Presented at: PETRA '21; June 29-July 2, 2021; Corfu, Greece. [doi: 10.1145/3453892.3461323]
31. Vergeer P, van Schaik Y, Sjerps M. Measuring calibration of likelihood-ratio systems: a comparison of four metrics, including a new metric devPAV. Forensic Sci Int 2021 Apr;321:110722. [doi: 10.1016/j.forsciint.2021.110722] [Medline: 33684845]
32. Lagioia F, Rovatti R, Sartor G. Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. AI Soc 2022 Apr 28;38(2):459-478. [doi: 10.1007/s00146-022-01441-y]
33. Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y. How do fairness definitions fare?: examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019 Presented at: AIES '19; January 27-28, 2019; Honolulu, HI. [doi: 10.1145/3306618.3314248]
34. Park Y, Hu J, Singh M, Sylla I, Dankwa-Mullan I, Koski E, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open 2021 Apr 01;4(4):e213909 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.3909] [Medline: 33856478]
35. Xu J, Xiao Y, Wang WH, Ning Y, Shenkman EA, Bian J, et al. Algorithmic fairness in computational medicine. EBioMedicine 2022 Oct;84:104250 [FREE Full text] [doi: 10.1016/j.ebiom.2022.104250] [Medline: 36084616]
36. Tao G, Sun W, Han T, Fang C, Zhang X. RULER: discriminative and iterative adversarial training for deep neural network fairness. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2022 Presented at: ESEC/FSE '22; November 14-18, 2022; Singapore. [doi: 10.1145/3540250.3549169]
37. Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. Commun ACM 2020 Apr 20;63(5):82-89. [doi: 10.1145/3376898]
38. Yao S, Huang B. Beyond parity: fairness objectives for collaborative filtering. arXiv Preprint posted online March 24, 2017 [FREE Full text]
39. Zhang Y, Zhou L. Fairness assessment for artificial intelligence in financial industry. arXiv Preprint posted online December 16, 2019 [FREE Full text] [doi: 10.5260/chara.21.2.8]

40. Ghassami A, Khodadadian S, Kiyavash N. Fairness in supervised learning: an information theoretic approach. arXiv Preprint posted online January 13, 2018 [FREE Full text] [doi: 10.1109/isit.2018.8437807]
41. Malawski M. A note on equal treatment and symmetry of values. In: Nguyen NT, Kowalczyk R, Mercik J, Motylska-Kuźma A, editors. Transactions on Computational Collective Intelligence XXXV. Berlin, Heidelberg: Springer; 2020.
42. Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, et al. Aequitas: a bias and fairness audit toolkit. arXiv Preprint posted online November 14, 2018 [FREE Full text] [doi: 10.48550/arXiv.1811.05577]
43. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 2019 Jul 1;63(4/5):4:1-:15. [doi: 10.1147/jrd.2019.2942287]
44. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging 2021 Sep;6(9):856-864 [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
45. Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, et al. Fairlearn: a toolkit for assessing and improving fairness in AI. Microsoft. 2020 Sep 22. URL: https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf [accessed 2023-12-05]
46. Ghosh A, Genuit L, Reagan M. Characterizing intersectional group fairness with worst-case comparisons. arXiv Preprint posted online January 05, 2021 [FREE Full text]
47. Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP. Fairness constraints: a flexible approach for fair classification. J Mach Learn Res 2019;20(75):1-42.
48. Chakraborti T, Patra A, Noble JA. Contrastive fairness in machine learning. IEEE Lett Comput Soc 2020 Jul 7;3(2):38-41. [doi: 10.1109/locs.2020.3007845]
49. Rosenfeld A, Richardson A. Explainability in human–agent systems. Auton Agent Multi-Agent Syst 2019 May 13;33:673-705. [doi: 10.1007/s10458-019-09408-y]
50. Narasimhan H, Cotter A, Gupta M, Wang S. Pairwise fairness for ranking and regression. Proc AAAI Conf Artif Intell 2020 Apr 03;34(04):5248-5255. [doi: 10.1609/aaai.v34i04.5970]
51. Kaur D, Uslu S, Durresi A, Badve S, Dundar M. Trustworthy explainability acceptance: a new metric to measure the trustworthiness of interpretable ai medical diagnostic systems. In: Proceedings of the 15th International Conference on Complex, Intelligent and Software Intensive Systems. 2021 Presented at: CISIS-2021; July 1-3, 2021; Asan, Korea. [doi: 10.1007/978-3-030-79725-6_4]
52. Bucher M, Herbin S, Jurie F. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In: Proceedings of the Computer Vision – ECCV 2016. 2016 Presented at: ECCV 2016; October 11-14, 2016; Amsterdam, The Netherlands. [doi: 10.1007/978-3-319-46454-1_44]
53. Kynkäänniemi T, Karras T, Laine S, Lehtinen J, Aila T. Improved precision and recall metric for assessing generative models. arXiv Preprint posted online April 15, 2019 [FREE Full text]
54. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv Preprint posted online August 13, 2020 [FREE Full text] [doi: 10.48550/arXiv.2008.05756]
55. Zaki MM, Jena AB, Chandra A. Supporting value-based health care - aligning financial and legal accountability. N Engl J Med 2021 Sep 09;385(11):965-967. [doi: 10.1056/NEJMp2105625] [Medline: 34478249]
56. Blacklaws C. Algorithms: transparency and accountability. Philos Trans A Math Phys Eng Sci 2018 Sep 13;376(2128):20170351. [doi: 10.1098/rsta.2017.0351] [Medline: 30082299]
57. Kim B, Park J, Suh J. Transparency and accountability in AI decision support: explaining and visualizing convolutional neural networks for text information. Decis Support Syst 2020 Jul;134:113302. [doi: 10.1016/j.dss.2020.113302]
58. Dubberley S, Murray D, Koenig A. Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation, and Accountability. Oxford, UK: Oxford University Press; 2020.
59. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019 Presented at: FAT* '19; January 29-31, 2019; Atlanta, GA. [doi: 10.1145/3287560.3287596]
60. King J. The instrumental value of legal accountability. In: Accountability in the Contemporary Constitution. Oxford, UK: Oxford University Press; 2013.
61. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019 Nov 04;1:501-507. [doi: 10.1038/s42256-019-0114-4]
62. Kass NE, Faden RR. Ethics and learning health care: the essential roles of engagement, transparency, and accountability. Learn Health Syst 2018 Sep 18;2(4):e10066 [FREE Full text] [doi: 10.1002/lrh2.10066] [Medline: 31245590]
63. Wright D. A framework for the ethical impact assessment of information technology. Ethics Inf Technol 2010 Jul 8;13:199-226. [doi: 10.1007/s10676-010-9242-6]
64. Arnold T, Kasenberg D, Scheutz M. Value alignment or misalignment – what will keep systems accountable? Association for the Advancement of Artificial Intelligence. 2017. URL: https://hrilab.tufts.edu/publications/arnoldetal17aiethics.pdf [accessed 2023-12-05]

65. Iyer R, Li Y, Li H, Lewis M, Sundar R, Sycara K. Transparency and explanation in deep reinforcement learning neural networks. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018 Presented at: AIES '18; February 2-3, 2018; New Orleans, LA. [doi: 10.1145/3278721.3278776]
66. Fukuda‐Parr S, Gibbons E. Emerging consensus on 'ethical AI': human rights critique of stakeholder guidelines. Glob Policy 2021 Jun 19;12(S6):32-44. [doi: 10.1111/1758-5899.12965]
67. Wachter S, Mittelstadt B, Floridi L. Transparent, explainable, and accountable AI for robotics. Sci Robot 2017 May 31;2(6):eaan6080. [doi: 10.1126/scirobotics.aan6080] [Medline: 33157874]
68. Ozga J. The politics of accountability. J Educ Change 2020;21:19-35. [doi: 10.1007/s10833-019-09354-2]
69. Ko RK, Kirchberg M, Lee BS. From system-centric to data-centric logging - accountability, trust and security in cloud computing. In: Proceedings of the Defense Science Research Conference and Expo. 2011 Presented at: DSR 2011; August 3-5, 2011; Singapore. [doi: 10.1109/dsr.2011.6026885]
70. Raji I, Smart A, White R, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020 Presented at: FAT* '20; January 27-30, 2020; Barcelona, Spain. [doi: 10.1145/3351095.3372873]
71. Nushi B, Kamar E, Horvitz E. Towards accountable AI: hybrid human-machine analyses for characterizing system failure. Proc AAAI Conf Hum Comput Crowdsourc 2018 Jun 15;6(1):126-135. [doi: 10.1609/hcomp.v6i1.13337]
72. Vesnic-Alujevic L, Nascimento S, Pólvora A. Societal and ethical impacts of artificial intelligence: critical notes on European policy frameworks. Telecommun Policy 2020 Jul;44(6):101961. [doi: 10.1016/j.telpol.2020.101961]
73. Kerikmäe T, Pärn-Lee E. Legal dilemmas of Estonian artificial intelligence strategy: in between of e-society and global race. AI Soc 2020 Jul 01;36:561-572. [doi: 10.1007/s00146-020-01009-8]
74. Reich MR. The core roles of transparency and accountability in the governance of global health public-private partnerships. Health Syst Reform 2018;4(3):239-248. [doi: 10.1080/23288604.2018.1465880] [Medline: 30207904]
75. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016 Jun;9:77-82 [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
76. Laacke S, Mueller R, Schomerus G, Salloch S. Artificial intelligence, social media and depression. A new concept of health-related digital autonomy. Am J Bioeth 2021 Jul;21(7):4-20. [doi: 10.1080/15265161.2020.1863515] [Medline: 33393864]
77. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013 Oct;46(5):830-836 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.010] [Medline: 23820016]
78. Crawley AW, Divi N, Smolinski MS. Using timeliness metrics to track progress and identify gaps in disease surveillance. Health Secur 2021;19(3):309-317. [doi: 10.1089/hs.2020.0139] [Medline: 33891487]
79. Zhai C, Cohen WW, Lafferty J. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. ACM SIGIR Forum 2015 Jun 23;49(1):2-9. [doi: 10.1145/2795403.2795405]
80. Weiss DJ, Nelson A, Gibson HS, Temperley W, Peedell S, Lieber A, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature 2018 Jan 18;553(7688):333-336 [FREE Full text] [doi: 10.1038/nature25181] [Medline: 29320477]
81. Burgess K, Hart D, Elsayed A, Cerny T, Bures M, Tisnovsky P. Visualizing architectural evolution via provenance tracking: a systematic review. In: Proceedings of the Conference on Research in Adaptive and Convergent Systems. 2022 Presented at: RACS '22; October 3-6, 2022; Virtual event. [doi: 10.1145/3538641.3561493]
82. Diakopoulos N, Koliska M. Algorithmic transparency in the news media. Digit J 2016 Jul 27;5(7):809-828. [doi: 10.1080/21670811.2016.1208053]
83. Stellefson M, Paige SR, Chaney BH, Chaney JD. Evolving role of social media in health promotion: updated responsibilities for health education specialists. Int J Environ Res Public Health 2020 Feb 12;17(4):1153 [FREE Full text] [doi: 10.3390/ijerph17041153] [Medline: 32059561]
84. Valko M, Hauskrecht M. Feature importance analysis for patient management decisions. Stud Health Technol Inform 2010;160(Pt 2):861-865 [FREE Full text] [Medline: 20841808]
85. Lipton ZC. The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 2018 Jun 01;16(3):31-57. [doi: 10.1145/3236386.3241340]
86. Slack D, Friedler SA, Scheidegger C, Dutta Roy C. Assessing the local interpretability of machine learning models. arXiv Preprint posted online February 9, 2019 [FREE Full text]
87. Stepin I, Alonso JM, Catala A, Pereira-Farina M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access 2021 Jan 13;9:11974-12001. [doi: 10.1109/access.2021.3051315]
88. Bertino E, Merrill S, Nesen A, Utz C. Redefining data transparency: a multidimensional approach. Computer 2019 Jan;52(1):16-26. [doi: 10.1109/MC.2018.2890190]
89. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
90. Li Q. Overview of data visualization. In: Embodying Data. Singapore: Springer; Jun 20, 2020.

91. Azeroual O, Saake G, Schallehn E. Analyzing data quality issues in research information systems via data profiling. Int J Inf Manage 2018 Aug;41:50-56. [doi: 10.1016/j.ijinfomgt.2018.02.007]
92. Tang M, Shao S, Yang W, Liang Y, Yu Y, Saha B, et al. SAC: a system for big data lineage tracking. In: Proceedings of the IEEE 35th International Conference on Data Engineering (ICDE). 2019 Presented at: ICDE 2019; April 8-11, 2019; Macao, China. [doi: 10.1109/icde.2019.00215]
93. Leslie D. Understanding artificial intelligence ethics and safety. arXiv Preprint posted online June 11, 2019 [FREE Full text]
94. Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interact Intell Syst 2020 Oct 16;10(4):1-31. [doi: 10.1145/3419764]
95. Brundage M, Avin S, Wang J, Belfield H, Krueger D, Hadfield G, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv Preprint posted online April 15, 2020 [FREE Full text] [doi: 10.48550/ARXIV.2004.07213]
96. Janssen M, Hartog M, Matheus R, Yi Ding A, Kuk G. Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. Soc Sci Comput Rev 2020 Dec 28;40(2):478-493. [doi: 10.1177/0894439320980118]
97. Paredes JN, Teze JC, Martinez MV, Simari GI. The HEIC application framework for implementing XAI-based socio-technical systems. Online Soc Netw Media 2022 Nov;32:100239. [doi: 10.1016/j.osnem.2022.100239]
98. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020 Nov 30;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
99. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020 Jun;58:82-115. [doi: 10.1016/j.inffus.2019.12.012]
100. Xiong Z, Cui Y, Liu Z, Zhao Y, Hu M, Hu J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Comput Materials Sci 2020 Jan;171:109203. [doi: 10.1016/j.commatsci.2019.109203]
101. Zafar MR, Khan N. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach Learn Knowl Extr 2021 Jun 30;3(3):525-541. [doi: 10.3390/make3030027]
102. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv Preprint posted online May 2, 2017 [FREE Full text]
103. Sokol K, Flach P. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In: Proceedings of the AAAI Workshop on Artificial Intelligence Safety, SafeAI 2019. 2019 Presented at: SafeAI 2019; January 27, 2019; Honolulu, HI.
104. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021 Nov;32(11):4793-4813. [doi: 10.1109/TNNLS.2020.3027314] [Medline: 33079674]
105. Vig J. A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2019 Presented at: ACL 2019; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-3007]
106. Fan CY, Chang PC, Lin JJ, Hsieh JC. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput 2011 Jan;11(1):632-644. [doi: 10.1016/j.asoc.2009.12.023]
107. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A 2019 Oct 29;116(44):22071-22080 [FREE Full text] [doi: 10.1073/pnas.1900654116] [Medline: 31619572]
108. Leikas J, Koivisto R, Gotcheva N. Ethical framework for designing autonomous intelligent systems. J Open Innov Technol Mark Complex 2019 Mar;5(1):18. [doi: 10.3390/joitmc5010018]
109. Latonero M. Governing artificial intelligence: upholding human rights and dignity. Data & Society. 2018. URL: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf [accessed 2023-12-05]
110. Aiello AE, Renson A, Zivich PN. Social media- and internet-based disease surveillance for public health. Annu Rev Public Health 2020 Apr 02;41:101-118 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094402] [Medline: 31905322]
111. Olteanu A, Castillo C, Diaz F, Kıcıman E. Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data 2019;2:13 [FREE Full text] [doi: 10.3389/fdata.2019.00013] [Medline: 33693336]
112. Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018 Presented at: AIES '18; February 2-3, 2018; New Orleans, LA. [doi: 10.1145/3278721.3278729]
113. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD '15; August 10-13, 2015; Sydney, Australia. [doi: 10.1145/2783258.2783311]

114. Mendes R, Cunha M, Vilela JP, Beresford AR. Enhancing user privacy in mobile devices through prediction of privacy preferences. In: Proceedings of the 27th European Symposium on Research in Computer Security. 2022 Presented at: ESORICS 2022; September 26-30, 2022; Copenhagen, Denmark. [doi: 10.1007/978-3-031-17140-6_8]
115. Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: Proceedings of the IEEE Symposium on Security and Privacy (SP). 2016 Presented at: SP 2016; May 22-26, 2016; San Jose, CA. [doi: 10.1109/sp.2016.42]
116. Kazim E, Koshiyama AS. A high-level overview of AI ethics. Patterns (N Y) 2021 Sep 10;2(9):100314 [FREE Full text] [doi: 10.1016/j.patter.2021.100314] [Medline: 34553166]
117. Nebeker C, Parrish EM, Graham S. The AI-powered digital health sector: ethical and regulatory considerations when developing digital mental health tools for the older adult demographic. In: Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues. Cham, Switzerland: Springer; 2022:159-176.
118. Crockett MJ. Models of morality. Trends Cogn Sci 2013 Aug;17(8):363-366 [FREE Full text] [doi: 10.1016/j.tics.2013.06.005] [Medline: 23845564]
119. Colman AM. Game Theory and its Applications: In the Social and Biological Sciences. London, UK: Psychology Press; 1995.
120. Someh IA, Davern M, Breidbach C, Shanks G. Ethical issues in big data analytics: a stakeholder perspective. Commun Assoc Inf Syst 2019 May;44(34):718-747 [FREE Full text] [doi: 10.17705/1CAIS.04434]
121. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P T 2014 Jul;39(7):491-520 [FREE Full text] [Medline: 25083128]
122. Ponce SB, M Barry M, S Dizon D, S Katz M, Murphy M, Teplinsky E, et al. Netiquette for social media engagement for oncology professionals. Future Oncol 2022 Mar;18(9):1133-1141 [FREE Full text] [doi: 10.2217/fon-2021-1366] [Medline: 35109663]
123. Drabiak K, Wolfson J. What should health care organizations do to reduce billing fraud and abuse? AMA J Ethics 2020 Mar 01;22(3):E221-E231 [FREE Full text] [doi: 10.1001/amajethics.2020.221] [Medline: 32220269]
124. Neville P, Waylen A. Social media and dentistry: some reflections on e-professionalism. Br Dent J 2015 Apr 24;218(8):475-478. [doi: 10.1038/sj.bdj.2015.294] [Medline: 25908363]
125. Ennis-O-Connor M, Mannion R. Social media networks and leadership ethics in healthcare. Healthc Manage Forum 2020 May;33(3):145-148. [doi: 10.1177/0840470419893773] [Medline: 31884833]
126. Garg T, Shrigiriwar A. Managing expectations: how to navigate legal and ethical boundaries in the era of social media. Clin Imaging 2021 Apr;72:175-177. [doi: 10.1016/j.clinimag.2020.11.005] [Medline: 33296827]
127. Kalkman S, Mostert M, Gerlinger C, van Delden JJ, van Thiel GJ. Responsible data sharing in international health research: a systematic review of principles and norms. BMC Med Ethics 2019 Mar 28;20(1):21 [FREE Full text] [doi: 10.1186/s12910-019-0359-9] [Medline: 30922290]
128. Sharma S. Data Privacy and GDPR Handbook. Hoboken, NJ: John Wiley & Sons; 2019.
129. Leidner JL, Plachouras V. Ethical by design: ethics best practices for natural language processing. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. 2017 Presented at: EthNLP@EACL; April 4, 2017; Valencia, Spain. [doi: 10.18653/v1/w17-1604]
130. Guttman N. Ethical issues in health promotion and communication interventions. Oxford Research Encyclopedias Communication. 2017 Feb 27. URL: https://www.academia.edu/100398082/Ethical_Issues_in_Health_Promotion_and_Communication_Interventions [accessed 2023-12-05]
131. Denecke K, Bamidis P, Bond C, Gabarron E, Househ M, Lau AY, et al. Ethical issues of social media usage in healthcare. Yearb Med Inform 2015 Aug 13;10(1):137-147 [FREE Full text] [doi: 10.15265/IY-2015-001] [Medline: 26293861]
132. Gagnon K, Sabus C. Professionalism in a digital age: opportunities and considerations for using social media in health care. Phys Ther 2015 Mar;95(3):406-414. [doi: 10.2522/ptj.20130227] [Medline: 24903111]
133. Bhatia-Lin A, Boon-Dooley A, Roberts MK, Pronai C, Fisher D, Parker L, et al. Ethical and regulatory considerations for using social media platforms to locate and track research participants. Am J Bioeth 2019 Jun;19(6):47-61 [FREE Full text] [doi: 10.1080/15265161.2019.1602176] [Medline: 31135323]
134. Davis K, Patterson D. Ethics of Big Data. Sebastopol, CA: O'Reilly Media; Sep 2012.
135. Livingston JD, Milne T, Fang ML, Amari E. The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review. Addiction 2012 Jan;107(1):39-50 [FREE Full text] [doi: 10.1111/j.1360-0443.2011.03601.x] [Medline: 21815959]
136. Jakesch M, Buçinca Z, Amershi S, Olteanu A. How different groups prioritize ethical values for responsible AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022 Presented at: FAccT '22; June 21-24, 2022; Seoul, Republic of Korea. [doi: 10.1145/3531146.3533097]
137. Pastaltzidis I, Dimitriou N, Quezada-Tavarez K, Aidinlis S, Marquenie T, Gurzawska A, et al. Data augmentation for fairness-aware machine learning: preventing algorithmic bias in law enforcement systems. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022 Presented at: FAccT '22; June 21-24, 2022; Seoul, Republic of Korea. [doi: 10.1145/3531146.3534644]

138. Keshk M, Moustafa N, Sitnikova E, Turnbull B. Privacy-preserving big data analytics for cyber-physical systems. Wireless Netw 2018 Dec 20;28(3):1241-1249. [doi: 10.1007/s11276-018-01912-5]
139. Kayaalp M. Patient privacy in the era of big data. Balkan Med J 2018 Jan 20;35(1):8-17 [FREE Full text] [doi: 10.4274/balkanmedj.2017.0966] [Medline: 28903886]
140. Enarsson T, Enqvist L, Naarttijärvi M. Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. Inf Commun Technol Law 2021 Jul 27;31(1):123-153. [doi: 10.1080/13600834.2021.1958860]
141. Umbrello S, van de Poel I. Mapping value sensitive design onto AI for social good principles. AI Ethics 2021 Feb 01;1(3):283-296 [FREE Full text] [doi: 10.1007/s43681-021-00038-3] [Medline: 34790942]
142. Hossin M, Sulaiman MN, Mustapha A, Mustapha N, Rahmat RW. A hybrid evaluation metric for optimizing classifier. In: Proceedings of the 3rd Conference on Data Mining and Optimization (DMO). 2011 Presented at: DMO 2011; June 28-29, 2011; Putrajaya, Malaysia. [doi: 10.1109/dmo.2011.5976522]
143. Nguyen AT, Raff E, Nicholas C, Holt J. Leveraging uncertainty for improved static malware detection under extreme false positive constraints. arXiv Preprint posted online August 9, 2021 [FREE Full text] [doi: 10.48550/arXiv.2108.04081]
144. Jadon S. A survey of loss functions for semantic segmentation. In: Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. 2020 Presented at: CIBCB 2020; October 27-29, 2020; Via del Mar, Chile. [doi: 10.1109/cibcb48159.2020.9277638]
145. Hansen E. HIPAA (Health Insurance Portability and Accountability Act) rules: federal and state enforcement. Med Interface 1997 Aug;10(8):96-8, 101. [Medline: 10169779]
146. Grajales FJ3, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 2014 Feb 11;16(2):e13 [FREE Full text] [doi: 10.2196/jmir.2912] [Medline: 24518354]
147. Chen J, Wang Y. Social media use for health purposes: systematic review. J Med Internet Res 2021 May 12;23(5):e17917 [FREE Full text] [doi: 10.2196/17917] [Medline: 33978589]

Abbreviations

AI: artificial intelligence
AMIA: American Medical Informatics Association
FATE: fairness, accountability, transparency, and ethics
GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LIME: local interpretable model-agnostic explanations
ML: machine learning
NLP: natural language processing
PDP: partial dependence plot
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
RQ: research question
SHAP: Shapley additive explanations
SMP: social media platform

Edited by A Castonguay; submitted 18.06.23; peer-reviewed by G Randhawa, D Valdes, M Arab-Zozani; comments to author 28.10.23; revised version received 21.12.23; accepted 15.02.24; published 03.04.24.

Please cite as:
Singhal A, Neveditsin N, Tanveer H, Mago V
Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review
JMIR Med Inform 2024;12:e50048
URL: https://medinform.jmir.org/2024/1/e50048
doi:10.2196/50048
PMID:38568737

Review

Evaluating the Prevalence of Burnout Among Health Care Professionals Related to Electronic Health Record Use: Systematic Review and Meta-Analysis

Yuxuan Wu[1], MMed; Mingyue Wu[2], MSc; Changyu Wang[3], BSc; Jie Lin[4*], MSN; Jialin Liu[1,2*], MD; Siru Liu[5], PhD

[1]Department of Medical Informatics, West China Hospital, Sichuan University, Chengdu, China

[2]Information Center, West China Hospital, Sichuan University, Chengdu, China

[3]West China College of Stomatology, Sichuan University, Chengdu, China

[4]Department of Oral Implantology, West China Hospital of Stomatology, Sichuan University, Chengdu, China

[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[*]these authors contributed equally

Corresponding Author:
Jialin Liu, MD
Information Center
West China Hospital
Sichuan University
37 Guoxue Road
Chengdu, 610041
China
Phone: 86 28 85422306
Fax: 86 28 85582944
Email: Dljl8@163.com

Abstract

Background: Burnout among health care professionals is a significant concern, with detrimental effects on health care service quality and patient outcomes. The use of the electronic health record (EHR) system has been identified as a significant contributor to burnout among health care professionals.

Objective: This systematic review and meta-analysis aims to assess the prevalence of burnout among health care professionals associated with the use of the EHR system, thereby providing evidence to improve health information systems and develop strategies to measure and mitigate burnout.

Methods: We conducted a comprehensive search of the PubMed, Embase, and Web of Science databases for English-language peer-reviewed articles published between January 1, 2009, and December 31, 2022. Two independent reviewers applied inclusion and exclusion criteria, and study quality was assessed using the Joanna Briggs Institute checklist and the Newcastle-Ottawa Scale. Meta-analyses were performed using R (version 4.1.3; R Foundation for Statistical Computing), with EndNote X7 (Clarivate) for reference management.

Results: The review included 32 cross-sectional studies and 5 case-control studies with a total of 66,556 participants, mainly physicians and registered nurses. The pooled prevalence of burnout among health care professionals in cross-sectional studies was 40.4% (95% CI 37.5%-43.2%). Case-control studies indicated a higher likelihood of burnout among health care professionals who spent more time on EHR-related tasks outside work (odds ratio 2.43, 95% CI 2.31-2.57).

Conclusions: The findings highlight the association between the increased use of the EHR system and burnout among health care professionals. Potential solutions include optimizing EHR systems, implementing automated dictation or note-taking, employing scribes to reduce documentation burden, and leveraging artificial intelligence to enhance EHR system efficiency and reduce the risk of burnout.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42021281173; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021281173

(JMIR Med Inform 2024;12:e54811) doi:10.2196/54811

KEYWORDS

clinical decision support system; electronic health record; electronic medical record; health information technology; alert fatigue; burnout; health care professionals; health care service; EHR; systematic review; meta-analysis; health information system; clinician burnout; health informatics

Introduction

The integration of electronic health record (EHR) into health care systems marks the beginning of a new era in medical information management, with significant potential benefits for patient care, clinical decision-making, and administrative efficiency [1,2]. EHR systems are central to the modern health care infrastructure [3]. Along with these benefits, however, the widespread adoption of EHR systems has raised concerns about the well-being of health care professionals [4,5]. Unintended consequences, such as burnout among health care professionals, technology-related errors, and increased safety risks, have been associated with EHR use [4,6,7]. In addition, a notable part of the problems with EHR systems in the United States is the need to provide additional documentation for insurance companies [8].

Within the realm of EHR use, burnout among health care professionals, characterized by emotional exhaustion, depersonalization, and a diminished sense of personal accomplishment, has emerged as a critical concern [9,10]. Burnout among health care professionals has become a pressing public health concern [11-13]. Some studies have reported an average burnout prevalence of 44% [2], with rates exceeding 80% in some specific settings and departments [4,5] such as primary care and emergency departments. This pervasive problem affects not only health care professionals but also patients, with negative consequences such as reduced quality of care and increased medical errors and psychological problems [14-17]. The estimated annual cost of burnout among health care professionals due to medical negligence and staff turnover exceeds US $4 billion [18].

The phenomenon of burnout among health care professionals goes beyond individual distress and has significant implications for patient safety, quality of care, and overall health system performance [14,15,19]. Understanding the prevalence and underlying factors of EHR-related burnout among health care professionals is critical to developing effective interventions and policy adaptations. These interventions are essential to mitigate this burden and ensure the long-term sustainability of EHR implementation in health care [19,20]. The increase in EHR-related burnout among health care professionals reflects a multifaceted interplay of factors, including increased documentation requirements, cumbersome user interfaces, and the rapid pace of technological development [9,16,18].

This systematic review and meta-analysis aims to provide a comprehensive assessment of the existing literature on EHR-related provider burnout. It seeks to capture the full extent of burnout, identify its causes, and provide evidence-based support and recommendations to alleviate this pervasive problem. In addition, we hypothesize that specific features of EHR systems, such as user interface design or increased documentation requirements, may contribute to provider burnout. We hope that this work will serve as a guide for health care organizations, policy makers, and EHR developers in developing interventions and technological improvements that prioritize the well-being of health care professionals. In doing so, we can promote a sustainable and resilient health care system while harnessing the potential benefits of EHR systems to improve patient care.

Methods

Study Guidelines

We focused on studies that directly measured burnout, as it is often considered in existing research to be a distinct emotional state, separate from depression or anxiety. This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines [21] and was registered with PROSPERO (CRD42021281173). Details of the guidelines and registration can be found in Multimedia Appendices 1 and 2, respectively.

Definitions

Our definitions of burnout were based on the Maslach Burnout Inventory-Human Services Survey instrument (MBI-HSS) [13,22], which characterizes burnout with high emotional exhaustion as a score ≥27, high depersonalization as >10, and low personal accomplishment as <33. Across the included studies, burnout was defined inconsistently, with definitions ranging from any one of the 3 items to all 3 items. In cases where the same study examined multiple definitions of burnout, we used the most common definition (high emotional exhaustion, high depersonalization, and low personal accomplishment) for the meta-analysis. For alternative definitions, such as those from the Stanford Physician Wellness Survey [23] or mini-Z [24], only outcomes explicitly described as burnout were documented. We categorized studies according to the measurement tool and definition of burnout.

Search Strategy

We systematically searched PubMed, Embase, and Web of Science to identify relevant peer-reviewed English language studies published between January 1, 2009, and December 31, 2022. We used several search terms to capture EHR systems, including “electronic health record” and its abbreviation “EHR,” as well as “electronic medical record (EMR)” and “computerized physician order entry (CPOE).” To capture the phenomenon of burnout, we used terms such as “burnout,” “alert fatigue,” and “exhaustion.” In defining our study participants, we considered a spectrum of health care professionals, including “doctor,” “clinician,” “physician,” “surgeon,” “medical staff,” and “health care provider.” On June 30, 2023, the researchers conducted a literature search in databases such as PubMed, Embase, and Web of Science, following the previously established search strategy. No papers were found that met the inclusion criteria for this review.

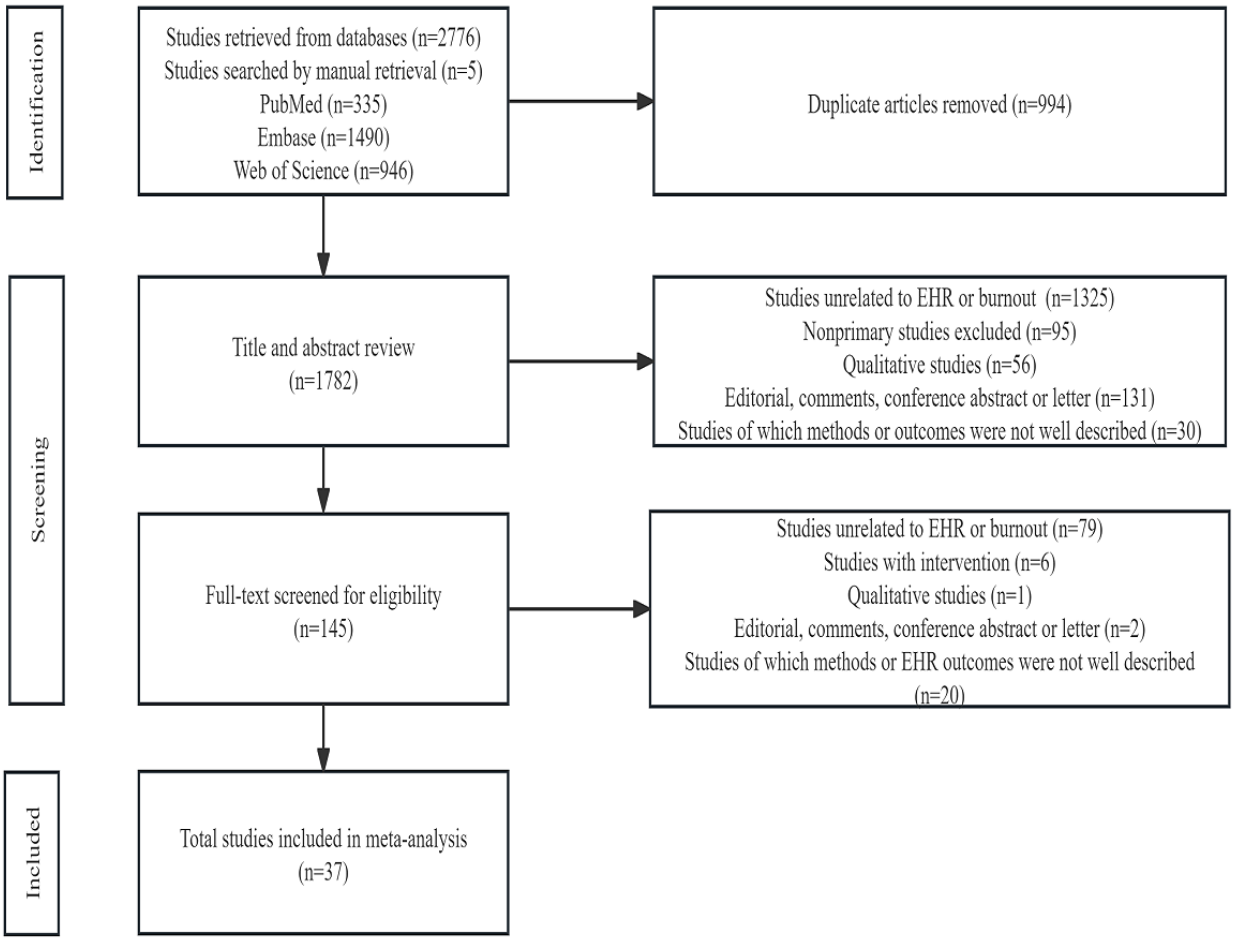
The terms were combined using Boolean logic and then filtered by publication date and language (English). A full description of the search strategy can be found in Multimedia Appendix 3. In addition, we carefully examined the references of each article and manually added 5 relevant references to our review list. Duplicate studies were systematically excluded from consideration.

Inclusion and Exclusion Criteria

Figure 1 shows the search and selection process. We applied strict inclusion and exclusion criteria to identify original and observational studies relevant to our research objectives. We included studies that examined general EHR use or specific supporting systems such as computerized physician order entry. We focused on studies that directly assessed burnout among health care professionals and individual psychological responses to EHR systems. Our review included the following types of research: cohort studies, case-control studies, and cross-sectional studies. EHR-related burnout was assessed using validated tools such as the MBI-HSS, the mini-Z, or other similar measures. The following publication types were excluded: abstracts, editorials, letters, reviews, commentaries, guidelines, and studies by non–health care professionals. In addition, studies were excluded if the necessary data could not be obtained from the corresponding author. We also excluded studies that repeated data already published in the literature.

Two reviewers independently screened all titles and abstracts to assess for relevance. Full texts of articles identified for further review were then assessed against the inclusion criteria. In cases of disagreement about the study eligibility of studies, a third reviewer was consulted for resolution.

Figure 1. Flowchart of study selection. EHR: electronic health record.



Data Extraction and Synthesis

For the included studies, we extracted relevant information including study design, geographical region, study duration, medical specialties involved, sample size, and relevant outcomes. The main outcomes included the prevalence of burnout in cross-sectional studies, the odds ratio (OR) along with its 95% CI in case-control studies, and the factors influencing burnout. We also documented the specific tools or measures used to assess these outcomes.

Risk of Bias Assessment

Two reviewers assessed the integrity, confirmability, and quality of the cross-sectional studies using the Joanna Briggs Institute (JBI) checklist [25] and the Newcastle-Ottawa Scale (NOS) [26] for the included case-control studies. Details of these assessments can be found in Multimedia Appendices 4-6, respectively.

Statistical Analysis

Meta-analysis was performed using R software (version 4.1.3; R Foundation for Statistical Computing). Heterogeneity was

calculated using the Cochran *Q* test, and statistical significance was set at $P<.05$. If there was no statistical heterogeneity ($I^2<50\%$), the fixed-effects model was used to pool results; otherwise ($I^2>50\%$) the random-effects model was used [27,28]. We grouped the main outcomes according to the predictor and moderator factors described by the participants and derived from the outcome reports. Continuous variables were summarized using the mean and standardized mean difference, whereas rates were extracted for categorical variables. For cross-sectional studies, the effect size measure was the prevalence of burnout and its corresponding 95% CI. For case-control studies, the effect of EHR was assessed using the pooled OR and its 95% CI. Publication bias was analyzed using the Egger test [29] and the trim-fill funnel plot. A sensitivity analysis was performed for each omitted method to determine the robustness and reliability of the results.

Results

Characteristics of the Included Studies

After reviewing a total of 2776 studies, 37 were selected for inclusion in our analysis (Figure 1) according to the predefined criteria and after elimination of duplicates. The baseline characteristics of the selected studies are summarized in Tables 1 and 2. For further details see Multimedia Appendix 7 [6,30-60].

The studies included in our review covered the period from 2009 to 2022 and included regions in both Canada and the United States. They involved a total of 66,556 health care professionals. The sample sizes of these studies varied widely from 84 to 25,018 participants, and the response rates ranged from 8.9% to 73.0%.

The primary measure used to assess burnout in the majority of studies was the MBI-HSS, which was used in 17 of 37 studies (46%). In addition, the mini-Z scale was used in 10 studies (27%). Notably, 2 studies using the MBI-HSS used cutoff definitions for burnout subcomponents that followed the standardized recommendations of the MBI-HSS.

Table 1. Characteristics of the cross-sectional studies.

Author	Data collection	Region	Participants	Sample (total)	Burnout cases	Burnout prevalence (%)
Tawfik et al [30]	2011	United States	Physicians and other clinician staff	6560	3586	54.66
Shanafelt et al [31]	2014	United States	Physicians	1934	517	26.73
Tawfik et al [32]	2015	United States	Physicians and other clinician staff	1165	624	53.56
Olson et al [33]	2016	United States	Physicians	282	127	45.04
Tai-Seale et al [34]	2016	United States	Physicians	107	41	38.32
Apaydin et al [35]	2016	United States	Physicians and other clinician staff	110	44	40
Livaudais et al [36]	2016	United States	Physicians and other clinician staff	557	267	47.94
Tran et al [37]	2017	United States	Physicians and other clinician staff	1792	465	25.95
Marckini et al [38]	2017	Canada and United States	Physicians	919	331	36.02
Gardner et al [39]	2017	United States	Physicians	208	51	24.52
Hilliard et al [40]	2017	United States	Physicians and other clinician staff	422	116	27.49
Higgins et al [41]	2017	United States	Residents	116	62	53.45
Czernik et al [42]	2017	United States	Residents	163	81	49.69
Hauer et al [43]	2018	United States	Physicians	122	44	36.07
Gajra et al [44]	2018	United States	Physicians	2468	539	21.84
Adler-Milstein et al [45]	2018	United States	Physicians	100	52	52
Somerson et al [46]	2018	United States	Residents	128	65	50.78
Melnick et al [47]	2018	United States	Physicians	203	78	38.42
Coleman et al [48]	2018	United States	Physicians	870	397	45.63
Abraham et al [49]	2018	United States	Nurses	368	134	36.41
Kondrich et al [50]	2018	Canada and United States	Physicians	872	360	41.28
Kroth et al [51]	2019	United States	Physicians and other clinician staff	856	276	32.24
Tajirian et al[6]	2019	Canada	Physicians and trainee	222	84	37.84
Mandeville et al [52]	2019	United States	Physicians and other clinician staff	396	100	25.25
Tiwari et al [53]	2019	United States	Physicians and other medical staff	15,505	5065	32.67
Sinha et al [54]	2019	United States	Physicians	103	41	39.81
Anderson et al [55]	2019	United States	Physicians and trainee	756	373	49.34
Nair et al [56]	2019	United States	Physicians	281	127	45.20
Jha et al [57]	2020	United States	Physicians and other medical staff	230	86	37.39
Esmaeilzadeh and Mirzaei [58]	2020	Iran	Physicians and other medical staff	416	206	49.52
Holzer et al [59]	2020	United States	Physicians and trainee	84	30	35.71
Wilkie et al [60]	2021	Canada	Physicians	457	106	23.19

Table 2. Characteristics of the case-control studies.

Author	Data collection	Participants	Region	Exposure	Sample (total)	Burnout cases	OR[a] (95% CI)
Eschenroeder et al [61]	2020	Physicians	United States	After-hours EHR[b] charting time per week >6 hours	25,018	7616	2.43 (2.30-2.57)
Sharp et al [62]	2019	Physician trainees	United States	Working hours per week >70 hours	502	159	2.80 (1.78-4.40)
Peccoralo et al [63]	2019	Clinical faculty	United States	Time spent on EHR outside work >90minutes	1346	385	1.90 (1.40-2.78)
Harris et al [64]	2017	Advanced practice registered nurses	United States	Insufficient time for EHR documentation	333	69	3.72 (1.78-7.80)
Robertson et al [65]	2015	Primary care workers	United States	Extra time spent on EHR per week >6 hours	585	216	2.90 (1.90-4.40)

[a]OR: odds ratio.

[b]EHR: electronic health record.

Quality of Included Studies

The quality of the cross-sectional studies was assessed using the JBI checklist. Of the cross-sectional studies reviewed, only 16 had a response rate of more than 50%. In addition, 24 studies provided a clear and precise description of their inclusion and exclusion criteria for participants. Additionally, 32 cross-sectional studies provided a detailed and thorough statistical analysis of their data and results.

We used the NOS to assess the risk of bias and the overall quality of the case-control studies. In particular, one study failed to clarify its selection criteria for the control group and comparability with the exposed group, which resulted in a high risk of selection bias. Furthermore, none of the 5 case-control studies reported information on the nonresponse population, indicating a high risk of nonresponse bias. Overall, the risk of bias in the case-control studies was assessed as moderate. A full breakdown of the quality assessment for each study can be found in Multimedia Appendices 4 [6,30-60] and 5 [61-65].

In our meta-analysis, we examined 37 studies that focused on identifying the prevalence of burnout associated with EHR use, involving a total of 66,556 health care professionals. The internal heterogeneity of 37 cross-sectional studies was evident in all included cross-sectional studies (I^2=98.3%). Using random-effects models, we calculated the combined overall prevalence of EHR-related burnout of 40.4% (95% CI 37.6%-43.2%). Subgroup analysis showed that studies using the MBI-HSS reported a higher pooled prevalence of burnout (41.4%) than those using the mini-Z (35.1%) but lower than those using other instruments (43.2%). However, these differences were not statistically significant (Figures 2 and 3).

Figure 2. Forest plot of the pooled prevalence of burnout among health care professionals across cross-sectional studies [6,30-60]. IV: inverse variance methods.

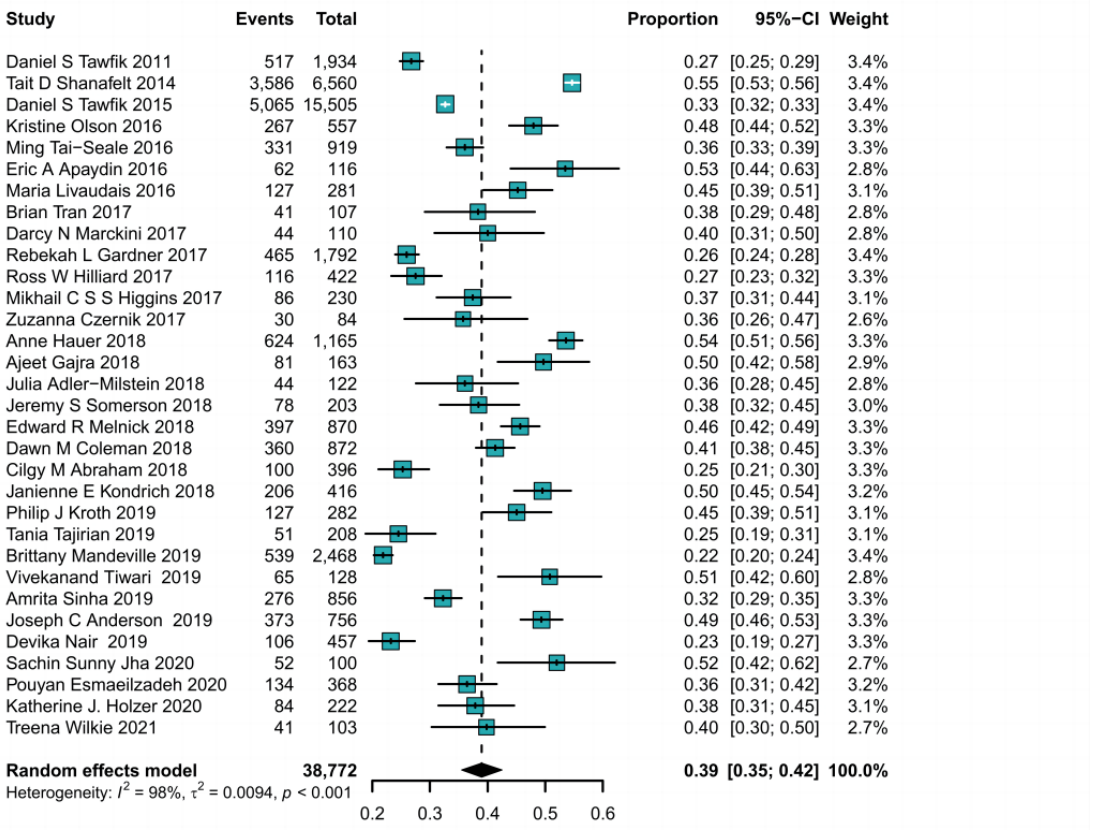
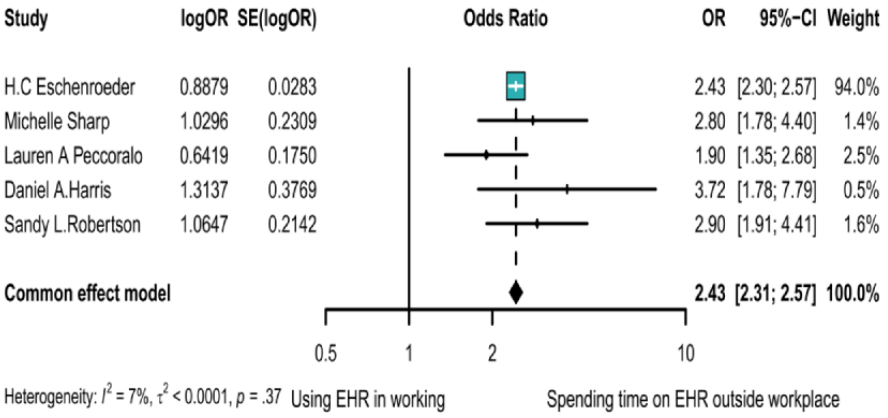


Figure 3. Measurement tool subgroup analysis of the pooled prevalence of burnout among health care professionals in cross-sectional studies [6,30-60]. IV: inverse variance methods; MBI: Maslach Burnout Inventory.



Publication Bias

The Egger test and the funnel plot were used to estimate the publication bias in the included studies (t=1.35, P=.18), indicating no significant publication bias. The distribution of the points in the funnel plot is symmetric. There was no statistical difference in publication bias. The results are available in Multimedia Appendices 8 and 9.

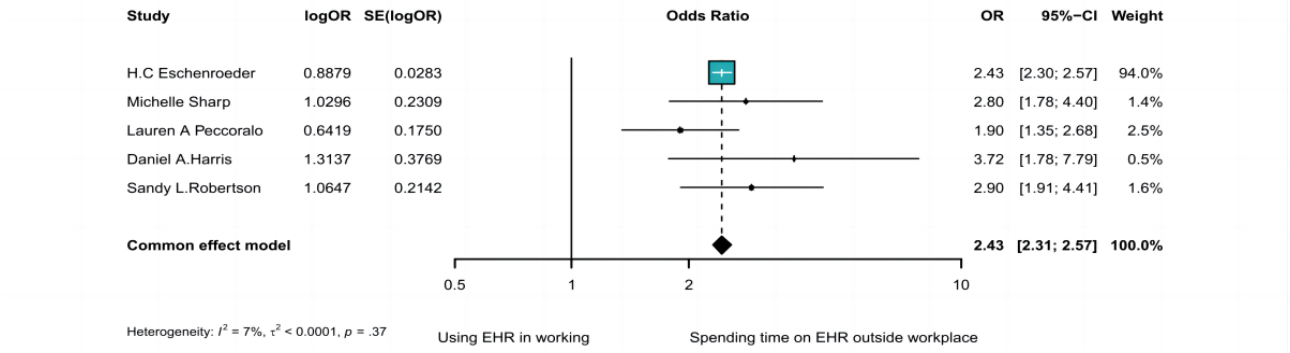
Sensitivity Analysis

Sensitivity analysis was performed using the individual omission method. The results showed that no single study had a significant effect on the pooled prevalence of burnout. The results of the sensitivity analysis indicated that the meta-analysis was robust.

The Association Between Time Spent on the EHR and Burnout

Data from 5 case-control studies with 27,784 participants were available for the meta-analysis of the time spent on EHR and burnout prevalence. There was no significant within-study heterogeneity (I^2=7.2%, P=.37), and a longer duration of EHR use was associated with a higher prevalence of burnout (OR 2.43, 95% CI 2.31-2.57) (Figure 4).

Figure 4. Forest plot of the association between the time spent using EHR and the risk of burnout [61-65]. EHR: electronic health record; IV: inverse variance methods; OR: odds ratio; SE: standard error of the TE; TE: take the logarithm of the effect value.



Main Causes of Burnout and Proposed Solutions

We have summarized the factors contributing to burnout among health care professionals in relation to EHR use in Table 3. Among these, challenges related to the design and availability of EHR systems were identified as the most significant contributors, as evidenced by 32 studies. Complaints from EHR users focused on several key issues: disruption to workflow [33]; cumbersome data entry (copy and paste) [59]; reduced direct communication with patients [38]; and annoyance with redundant, repetitive, or irrelevant alerts [52]. Poor EHR design has been shown to reduce work productivity and lead to prolonged EHR use [51,64]. This prolonged use has a negative impact on work-life balance of health care professionals and increasing burnout [42,57,65,66]. Workload factors, identified in 18 of the 32 studies, further exacerbate this problem. Specific aspects of workload that contribute to burnout include the number of hours worked per week [62-64], the frequency of night shifts [46,60], administrative documentation tasks [33,35,38,48,64], the volume of patient admissions [30,35,56], and the amount of information to manage in the EHR inbox [34,37,40]. Together, these factors exacerbate provider fatigue and increase the risk of burnout.

EHR usability, recognized as a contributing factor to burnout, relates to issues of accessibility and functionality of the system. This includes instances where the system is frequently unavailable due to maintenance, updates, or technical failures, as well as situations where the system is not user-friendly and requires excessive time to navigate and use effectively, potentially leading to burnout among health care professionals.

The factors contributing to burnout identified in the reviewed studies fall into 3 main categories: EHR use, work environment and organizational support, and the personal factors. Table 4 provides a summary of strategies to address these contributing factors. For example, the burden of medical clerical tasks imposed by EHR systems suggests the need to employ assistants or scribes to reduce the workload of health care professionals [31,67]. Evidence suggests that the EHR system itself can be improved by involving clinical staff in the design process [33], optimizing the user interface [39,64], minimizing the number of clicks required [52], and actively soliciting and incorporating user feedback [32]. In addition, some practitioners may not fully use health information technology in their roles and may be frustrated with EHR systems or similar systems [32]. To address this, health care organizations are advised to establish clear policies and procedures before implementing an EHR system and to provide ongoing health information technology education to reduce technology-related anxiety among users [32,52,68]. Finally, comprehensive and systematic initiatives are essential to effectively reduce burnout. Health care professionals are encouraged to work together to advocate for legislative and regulatory changes that ensure reasonable working hours, mandatory breaks, and safeguards against burnout [36,42,43,58,62].

Moreover, research also suggests that sociodemographic characteristics, interpersonal dynamics, and the work environment have a significant impact on the prevalence of burnout. In particular, factors such as being female, younger, and less experienced correlate with higher rates of burnout [34,48,55]. Conversely, high levels of satisfaction or positive perspectives on the use of EHR systems may reduce burnout [36,42,58].

Table 3. The influencing factors of burnout for studies.

Author	Design	Risk factors for burnout	Protective factors against burnout	Main EHR[a] factors influencing burnout
Tawfik et al [30]	Cross-sectional	NICU[b] with ≥10 weekly admissions, nursing care workload, and patient mortality	Burnout recognition education; implementation of burnout interventions at the individual and institutional level	Using EHR outside working or at home; time on using EHR
Shanafelt et al [31]	Cross-sectional	Using CPOE[c], female gender, emergency medicine, each additional hour per week	Assistant order entry; documentation support	Time spent on clerical tasks
Tawfik et al [32]	Cross-sectional	HIT[d] frustration, difficulty in falling asleep	Supplemental EHR training; scribes to assist documentation； team-based documentation and inbox management； automating data-entry tasks	Frustrated or stressed by EHR
Olson et al [33]	Cross-sectional	Poor control over workload, inefficient teamwork, lack of value alignment with leadership, and hectic-chaotic work atmosphere	Improve professional satisfaction; nonphysician order entry	Using EHR outside working or at home; insufficient documentation time
Tai-Seale et al [34]	Cross-sectional	Female gender and poor control over work schedule	Feeling highly valued; having good control over work schedule; working in a quiet or busy but reasonable environment; assist physician with email work; limit desktop medical work outside working hours (except in emergencies)	Using EHR outside working or at home; number of EHR system-generated in basket messages
Apaydin et al [35]	Cross-sectional	Managing unscheduled or same-day patients, lack of pharmacist support, administrative work, excessive overall workload, difficulty communicating with other professionals, inadequate care coordination, and answering patient emails	Interventions to facilitate provider-led quality improvement	Managing in-basket messages generated by EHR; responding to EHR alerts
Livaudais et al [36]	Cross-sectional	Negative perceptions of EHR	Perceiving positive effect of EHR in practice； technical support for EHR when using systems; EHR optimization program	Managing in-basket messages generated by EHR; poor EHR design; dealing with patient-call messages in systems
Tran et al [37]	Cross-sectional	Clinical full-time equivalents >0.9 and more incomplete messages in inbox	Perception positive attitudes about the effect of EHR or satisfied with EHR	Average additional 10 minutes spent on EHR after each visit; less efficient at completing EHR and inbox information
Marckini et al [38]	Cross-sectional	Female gender and dissatisfaction for clerical tasks	EHR optimization; improve physician efficiency; and job satisfaction	Managing in-basket messages generated by EHR; dissatisfaction with EHR
Gardner et al [39]	Cross-sectional	Primary care specialties, female gender, and reporting poor or marginal time for documentation	Perception positive attitudes about the effect of EHR or satisfied with EHR	Excessive data inputting in EHR; using EHR at home; frustrated with EHR
Hilliard et al [40]	Cross-sectional	High volume of patient call messages in the system and lack of control over workload	Copy and paste used in EHR documentation; assist with inbox tasks and create 2 administrative “desktops”	Using EHR outside working or at home; excessive data inputting in EHR; managing in-basket messages generated by EHR
Higgins et al [41]	Cross-sectional	Self-compassion, sleep disorder, lacking support from leaders, and poor control over schedules	Peer support, perceived appreciation and meaningfulness in work; maintaining values consistent with practice institution	Poor EHR usability; perception negative attitudes about the effect of EHR
Czernik et al [42]	Cross-sectional	Frustrated or stressed by EHR	Reducing the burden of documentation tasks; improving EHR usability; interventions to improve the EHR	Poor usability of EHR; information overload; degradation of medical documentation
Hauer et al [43]	Cross-sectional	Loss of practicing autonomy, female gender, frustrated with EHR, and increasing insurance and government regulation	Improve the functionality of EHR; enhance physician leadership and involvement; create a center for physician empowerment; create a physician health program	Using EHR outside workday

Author	Design	Risk factors for burnout	Protective factors against burnout	Main EHR[a] factors influencing burnout
Gajra et al [44]	Cross-sectional	Variable reimbursement models, interactions with payers, and increasing treating and caring demands	Use advanced practice providers; hire additional administrative staff; invest in information technology	Excessive data inputting in EHR; frustrated or stressed by EHR; using EHR outside workday
Adler-Milstein et al [45]	Cross-sectional	Poor self-rated EHR skills	Improve EHR design; scribe or team documentation; reduce documentation requirements	Using EHR outside working or at home; time spent on EHR; system-generated in-basket messages (>114) per week
Somerson et al [46]	Cross-sectional	Working >80 hours per week, verbal abuse from faculty, educational debt, “scut” work >10 hours per week	Nursing support; duty-hour restrictions; improve EHR functionality and efficiency; adequate, personalized training and support; adequate social work support	Time spent on EHR per week; used EHR >20 hours per week
Melnick et al [47]	Cross-sectional	Practice location (academic medical center) and medical specialty	Improve EHR usability	Using EHR outside working or at home; poor EHR usability
Coleman et al [48]	Cross-sectional	Work-related physical pain, work-home conflict, and younger age	Build personal resilience, enhance wellness; peer support; reduce administrative or EHR burden	Using EHR outside working or at home; increased EHR or documentation requirement
Abraham et al [49]	Cross-sectional	Intraorganizational factors	EHR with multifunctional; reduce high EHR workload; work with supportive colleagues; improve team communication	High EHR workload
Kondrich et al [50]	Cross-sectional	Feeling undervalued by patients, lacking superior support, little promotion chances, perceived unfair clinical working schedule, and nonacademic environment	Improve physician well-being	Feeling that the EHR detracts from patient care
Kroth et al [51]	Cross-sectional	Overall stress	Improve EHR design; clinician training; scribes to assist documentation; work at home boundaries; exercise, taking breaks	Information overloading; slow system response; excessive data inputting； fail to navigate quickly; note bloat; patient-clinician relationship interference; fear of missing something; billing oriented notes.
Tajirian et al [6]	Cross-sectional	Workflow issues	Reduce the administrative burden of EHR； improve EHR	Lower satisfaction and higher frustration with the EHR； poor intuitiveness and usability of EHR
Mandeville et al [52]	Cross-sectional	HIT-related stress and burnout and emergency medicine	Improved workflow	Daily frustration added by EHR； using EHR outside working or at home
Tiwari et al [53]	Cross-sectional	Lack of physical exercise and weekly working hours	Teamwork and working satisfaction; self-care training	Poor EHR usability; dissatisfaction with EHR
Sinha et al [54]	Cross-sectional	Interpersonal disengagement	Lower CLOC[e] ratio (total CLOC time to allocated appointment time); well-established personal resources	Using EHR outside working
Anderson et al [55]	Cross-sectional	Female gender, younger age, shorter practicing years, and having children at home	Taking 20 days or more of vacation time	Using EHR at home; ≥2-hour patient administration
Nair et al [56]	Cross-sectional	Working long hours, weekly number of nursing patients, practice environment, disinterested health systems, and dissatisfaction with remuneration	Caring for fewer patients per week	Using EHR outside working or at home; EHR requirements
Jha et al [57]	Cross-sectional	COVID-19 pandemic and in-house billing	Stay positive; improved EHR design	Documentation through EHR
Esmaeilzadeh and Mirzaei [58]	Cross-sectional	Less direct communication with patients, inadequate training for using HIT, and increasing computerization at work	Positive perceptions of EHR; more policy and legal interventions to ensure meaningful use of EHR	Poor EHR usability; time spent entering data
Holzer et al [59]	Cross-sectional	Receive COVID-19 patients	Using EHR to streamline clinical care activities; physician task relief	Using EHR outside work; increased EHR workload

Author	Design	Risk factors for burnout	Protective factors against burnout	Main EHR[a] factors influencing burnout
Wilkie et al [60]	Cross-sectional	High workload and insufficient resources	Good leadership; prioritize work-life balance	Poor EHR usability
Eschenroeder et al [61]	Case-control	Specialty	Organizational support for EHR	After-hours EHR charting time per week >6 hours; time-consuming data entry
Sharp et al [62]	Case-control	Working hours per week >70 hours	Report system to cover personal illness or emergency; access to mental health services; reduce EHR and clerical burden	>90 minutes on the EHR outside of the workday
Peccoralo et al [63]	Case-control	Clerical work time (>60 minutes/day) and poorer work-life integration	Reducing time spent on EHR and clerical tasks	Using EHR outside working (>90 minutes/day); EHR adds to daily work frustration
Harris et al [64]	Case-control	Insufficient time for documentation	Improve EHR usability; documentation practices optimization	Using EHR outside working or at home; EHR adding to daily frustration
Robertson et al [65]	Case-control	Dissatisfaction with work-life balance and female gender	EHR proficiency training	Extra time spent on EHR per week >6 hours

[a]EHR: electronic health record.

[b]NICU: neonatal intensive care unit.

[c]CPOE: computerized physician order entry.

[d]HIT: health information technology.

[e]CLOC: clinician logged-in outside clinic time.

Table 4. Proposed solutions for burnout mentioned.

Perspectives/solutions and suggestions	Measures
EHR[a]	
Improve EHR usability and performance	Enhance EHR user interface and design to reduce health care professionals to use
Institutions provide timely technical support during EHR use	Improving the effectiveness and efficiency of technological responses
Institutions should offer comprehensive training courses for EHR users	Ensure users master EHR skills to reduce burnout from technological issue
Working environment and organizational support	
Institutions introduce mechanisms for regular assessment of EHR efficacy	Regularly optimize and update the system based on user feedback
Establish a schedule, routine, and workflow	Design and optimize the workflow to ensure that the EHR aligns with the health care professional's actual work, reducing unnecessary steps and improving work efficiency
Enhance peer, managerial, and technical support	Provide appropriate human resources, such as medical assistants, scribes, and improving teamwork to distribute workload among health care professionals
Development of transparent policies and objectives	Establish clear policies and legislation to define the purpose, scope, and duration of EHR use, to delineate the responsibilities and obligations of health care professionals, and to reduce confusion and burnout
Personal	
Use of mental health resources and services	Counseling services and mindfulness meditation therapy help health care professionals better manage work stress and reduce their psychological distress
Encourage academic and career development	Plan career paths and training programs and create an environment for career development and learning

[a]EHR: electronic health record.

Discussion

Key Findings

This study explores the relationship between burnout and health care professionals. Our analysis revealed several key findings. First, the prevalence of burnout differs between assessment instruments, with the MBI-HSS indicating higher levels of burnout. However, this difference was not statistically significant. Second, there was a positive association between the average daily duration of EHR use and the risk of burnout. In particular, reducing the administrative burnout emerged as an effective strategy to reduce the risk of burnout [63]. Third, positive perceptions of the EHR and constructive work attitudes were correlated with the reduction in burnout.

The MBI-HSS is valued for its extensive validation and widespread acceptance as an essential tool for assessing burnout. Our findings suggest that the MBI-HSS may report higher rates of burnout due to several factors: sensitivity to burnout constructs—unlike self-report measures, which may rely predominantly on respondents' respondents' subjective feelings, the MBI-HSS comprehensively assesses burnout across multiple dimensions: emotional exhaustion, depersonalization, and personal accomplishment. This multidimensional assessment provides a nuanced perception of burnout, encompassing both its physical and psychological facets. These include the following: standardized cut-off scores—the MBI-HSS delineates specific cut-off scores for its dimensions, establishing clear criteria for identifying significant levels of burnout. This standardization promotes a consistent classification framework for burnout, which may contribute to the higher prevalence rates reported. Comprehensive assessment—the multidimensional approach of the MBI-HSS allows for a comprehensive assessment of burnout, including emotional exhaustion, depersonalization, and personal accomplishment. This thorough assessment is able to uncover more precise and detailed manifestations of burnout, thereby increasing detection rates. Benchmark for comparison—the MBI-HSS is often used as a benchmark for validating alternative burnout measures, and differences in results when compared with other instruments do not necessarily indicate a variance in prevalence. Rather, these differences underscore the accuracy of the MBI-HSS and the comprehensive scope of its assessment. The use of different instruments underlines the heterogeneity observed in our study results.

Solutions

This study demonstrates a robust relationship between workload, time spent using EHR, and burnout. Through a systematic review, we outline several pragmatic recommendations aimed at mitigating these problems.

Reduce Documentation and EHR Workload

A key strategy for alleviating workload concerns is to adopt a rational task allocation and effective teamwork model. Previous research highlights the effectiveness of this approach in reducing workload pressures [33,53]. By integrating medical assistants and scribes into the health care team, it is possible to distribute clerical tasks more evenly, thereby reducing the burden on health care professionals. This redistribution not only reduces workload but also increases overall operational efficiency [53,69,70]. In addition, the provision of targeted training is critical to improving teamwork dynamics, communication skills, and workflow efficiency. Such training efforts aim to cultivate a competent team capable of optimizing and streamlining workflow processes. The ultimate goal is to minimize documentation and EHR-related workloads, thereby making a significant contribution to reducing burnout among health care professionals [58,63].

Optimizing EHR and Training Courses

Continuous refinement of EHR systems through improved design, functionality, and integration of predesigned templates and phrases effectively increases system efficiency. The elimination of redundant steps and interactions further improves the user experience [32,71]. For example, customizing templates to include commonly used medical advice and alerts tailored to the specific needs of different departments significantly increases EHR efficiency [48,72]. Numerous studies have highlighted the critical role of improving user interaction with the EHR system. Developing a user-friendly interface that minimizes unnecessary clicks and reduces redundant and irrelevant data entry has been shown to significantly improve the user experience. Such improvements also significantly reduce the cognitive burden on health care professionals, resulting in a more streamlined and efficient health care delivery process [32,39,42]. In addition, comprehensive training and strong technical support are critical to improving the efficiency and effectiveness of EHR use. Systematic training aimed at promoting EHR proficiency among health care professionals can significantly improve operational efficiency and mitigate the effects of technology stress [46,58]. Research emphasizes the importance of training health care professionals to enhance EHR use and tailoring templates to specific clinical workflows.

Artificial Intelligence–Based Solutions

The integration of artificial intelligence (AI) into EHR systems represents a significant frontier for improvement. Innovations in machine learning, natural language processing (NLP), and large language models (LLMs) are poised to significantly increase the intelligence and automation capabilities of EHR systems [73,74]. Incorporating speech recognition and automated dictation or note-taking into hospital workflows can streamline the creation of medical documents, thereby increasing operational efficiency [75]. NLP is characterized by its ability to efficiently organize both unstructured and semistructured textual records, thereby facilitating a reduction in paperwork [76,77]. Recent research has highlighted the utility of LLMs, such as GPT-4, as powerful tools for medical documentation [78,79]. The use of technologies such as GPT-4 as a linguistic assistant or the use of intelligent templates can significantly speed up the medical documentation process for health care professionals, while improving the accuracy of documentation [79]. In addition, the researchers developed a data-driven method to generate recommendations for refining alert criteria through an explainable AI framework [80]. This advancement directly addresses the issue of overalerting in clinical decision support systems, which has been identified as a potential contributor to

burnout among health care professionals. By reducing unnecessary alerts, this approach promises to reduce the cognitive and operational workload of health care professionals, thereby improving both the quality of patient care and the work-life balance of health care staff. While AI technology could potentially help reduce burnout, it is important to recognize that the causes of burnout are complex and require further research.

Implications for Future Research

There is considerable evidence to support the need for comprehensive redesign of EHR systems to improve efficiency [32,51,53,81]. However, the literature reveals a paucity of published empirical research quantifying EHR limitations, user fatigue and burnout. While some studies have indirectly demonstrated the poor usability of EHR by measuring pupillary reflex and cognitive fatigue [82,83], claims of inefficiency are primarily based on subjective perceptions of users. Thus, there is a need for more studies that objectively assess usability and user experience. Future research should aim to quantitatively assess the usability of EHR systems and their impact on the physical and mental well-being of health care professionals.

Furthermore, the incorporation of AI, specifically LLMs, into EHR systems is an important future research direction to reduce burnout among health care professionals. Such research could include, but is not limited to, (1) reducing the amount of time health care professionals spend on nonclinical tasks by automating administrative tasks, including data entry, scheduling, and patient history taking; (2) using LLMs to efficiently generate and review medical documentation to ensure high quality and consistency of documentation while saving time; (3) improving the interpretability and transparency of clinical decision support to provide clinicians with trustworthy decision support to reduce their cognitive load; and (4) ensuring the ethical use of AI to guarantee that AI systems are used ethically and that algorithms are unbiased. The integration of AI into EHR systems must comply with strict privacy regulations to protect patient privacy [84]. Exploring the potential of AI could make a significant contribution to creating a more supportive and efficient health care ecosystem [73,79,85].

Limitations

This review has several limitations. First, it has a language bias by including only peer-reviewed literature published in English. This limitation may introduce information and selection bias by omitting non-English studies that may provide valuable insights or alternative viewpoints on the topic. Second, the internal heterogeneity of the included studies is remarkably high, with significant differences in methodology, participant demographics, and outcome measures between studies, which may bias the synthesis of findings. In addition, the geographical distribution of the selected studies is dominated by North American research, with only 1 study from Iran. This distribution may introduce regional bias, as health care practices and experiences in these areas may not accurately reflect global patterns.

In addition, the temporal scope of the study, covering the years 2020 to 2022, was significantly influenced by the COVID-19 pandemic. Data collected during this period may be subject to bias or inaccuracy due to the unprecedented impact of the pandemic on global health systems. Additionally, the pandemic introduced new stressors and challenges for health care professionals, which may have influenced the incidence and manifestation of their burnout. These factors should be carefully considered when interpreting the study results, as they may limit the generalizability and significance of the findings beyond the specific context and timeframe of the pandemic.

Conclusions

This review highlights the significant impact of the EHR and the workload of health care professionals on burnout and emphasizes the need for targeted solutions such as workflow optimization, improved training, and the use of medical scribes. It also identifies that the potential of AI to improve EHR efficiency is a promising direction. Despite these findings, there remains a critical need for empirical research to accurately quantify the challenges associated with EHR use and their impact on provider well-being. Future studies are encouraged to explore innovative solutions to foster a more supportive health care environment.

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Authors' Contributions

J Liu, J Lin, and SL conceived and designed the study. YW, SL, MW, J Liu, and J Lin developed the methods. YW, SL, CW, J Lin, and J Liu developed the search strategy. All authors participated in drafting the manuscript. All authors have read and approved the final article. There was no funding for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 98 KB - medinform_v12i1e54811_app1.pdf]

Multimedia Appendix 2
PROSPERO registration.
[PDF File (Adobe PDF File), 1901 KB - medinform_v12i1e54811_app2.pdf]

Multimedia Appendix 3
Search strategy.
[DOCX File , 13 KB - medinform_v12i1e54811_app3.docx]

Multimedia Appendix 4
Joanna Briggs Institute checklist for the cross-sectional studies included.
[PDF File (Adobe PDF File), 58 KB - medinform_v12i1e54811_app4.pdf]

Multimedia Appendix 5
NOS results for the case-control studies included.
[PDF File (Adobe PDF File), 74 KB - medinform_v12i1e54811_app5.pdf]

Multimedia Appendix 6
Joanna Briggs Institute Prevalence Critical Appraisal Tool.
[DOCX File , 15 KB - medinform_v12i1e54811_app6.docx]

Multimedia Appendix 7
Basic characteristics of the studies included.
[PDF File (Adobe PDF File), 70 KB - medinform_v12i1e54811_app7.pdf]

Multimedia Appendix 8
Funnel plot for the studies included.
[PNG File , 136 KB - medinform_v12i1e54811_app8.png]

Multimedia Appendix 9
The results of the publication bias test.
[PNG File , 45 KB - medinform_v12i1e54811_app9.png]

References

1. Aldosari B. Patients' safety in the era of EMR/EHR automation. Inform Med Unlocked 2017;9:230-233 [FREE Full text] [doi: 10.1016/j.imu.2017.10.001]
2. Gatiti P, Ndirangu E, Mwangi J, Mwanzu A, Ramadhani T. Enhancing healthcare quality in hospitals through electronic health records: a systematic review. J Health Inform Dev Ctries 2021;15(2):1-25 [FREE Full text]
3. Woldemariam MT, Jimma W. Adoption of electronic health record systems to enhance the quality of healthcare in low-income countries: a systematic review. BMJ Health Care Inform 2023 Jun;30(1):e100704 [FREE Full text] [doi: 10.1136/bmjhci-2022-100704] [Medline: 37308185]
4. Li C, Parpia C, Sriharan A, Keefe DT. Electronic medical record-related burnout in healthcare providers: a scoping review of outcomes and interventions. BMJ Open 2022 Aug 19;12(8):e060865 [FREE Full text] [doi: 10.1136/bmjopen-2022-060865] [Medline: 35985785]
5. Heponiemi T, Gluschkoff K, Vehko T, Kaihlanen AM, Saranto K, Nissinen S, et al. Electronic health record implementations and insufficient training endanger nurses' well-being: cross-sectional survey study. J Med Internet Res 2021 Dec 23;23(12):e27096 [FREE Full text] [doi: 10.2196/27096] [Medline: 34941546]
6. 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 Jul 15;22(7):e19274 [FREE Full text] [doi: 10.2196/19274] [Medline: 32673234]
7. Palojoki S, Saranto K, Reponen E, Skants N, Vakkuri A, Vuokko R. Classification of electronic health record-related patient safety incidents: development and validation study. JMIR Med Inform 2021 Aug 31;9(8):e30470 [FREE Full text] [doi: 10.2196/30470] [Medline: 34245558]
8. Tutty MA, Carlasare LE, Lloyd S, Sinsky CA. The complex case of EHRs: examining the factors impacting the EHR user experience. J Am Med Inform Assoc 2019 Jul 01;26(7):673-677 [FREE Full text] [doi: 10.1093/jamia/ocz021] [Medline: 30938754]
9. Jankovic I, Chen JH. Clinical decision support and implications for the clinician burnout crisis. Yearb Med Inform 2020 Aug;29(1):145-154 [FREE Full text] [doi: 10.1055/s-0040-1701986] [Medline: 32823308]

10. Thomas Craig KJ, Willis VC, Gruen D, Rhee K, Jackson GP. The burden of the digital environment: a systematic review on organization-directed workplace interventions to mitigate physician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):985-997 [FREE Full text] [doi: 10.1093/jamia/ocaa301] [Medline: 33463680]
11. Mohan V, Garrison C, Gold JA. Using a new model of electronic health record training to reduce physician burnout: a plan for action. JMIR Med Inform 2021 Sep 20;9(9):e29374 [FREE Full text] [doi: 10.2196/29374] [Medline: 34325400]
12. Melnick ER, Harry E, Sinsky CA, Dyrbye LN, Wang H, Trockel MT, et al. Perceived electronic health record usability as a predictor of task load and burnout among US physicians: mediation analysis. J Med Internet Res 2020 Dec 22;22(12):e23382 [FREE Full text] [doi: 10.2196/23382] [Medline: 33289493]
13. Mertz H. Electronic health record reform: an alternative response to physician burnout. Am J Med 2021;134(9):e498. [doi: 10.1016/j.amjmed.2021.04.022] [Medline: 34462089]
14. Dyrbye LN, Shanafelt TD, Sinsky CA, Cipriano PF, Bhatt J, Ommaya A, et al. Burnout among health care professionals: a call to explore and address this underrecognized threat to safe, high-quality care. NAM Perspectives 2017;7(7) [FREE Full text] [doi: 10.31478/201707b]
15. Kumar S. Burnout and doctors: prevalence, prevention and intervention. Healthcare (Basel) 2016;4(3):37 [FREE Full text] [doi: 10.3390/healthcare4030037] [Medline: 27417625]
16. Gesner E, Gazarian P, Dykes P. The burden and burnout in documenting patient care: an integrative literature review. Stud Health Technol Inform 2019;264:1194-1198 [FREE Full text] [doi: 10.3233/SHTI190415] [Medline: 31438114]
17. Kang C, Sarkar IN. Interventions to reduce electronic health record-related burnout: a systematic review. Appl Clin Inform 2024;15(1):10-25 [FREE Full text] [doi: 10.1055/a-2203-3787] [Medline: 37923381]
18. Johnson KB, Neuss MJ, Detmer DE. Electronic health records and clinician burnout: a story of three eras. J Am Med Inform Assoc 2021;28(5):967-973 [FREE Full text] [doi: 10.1093/jamia/ocaa274] [Medline: 33367815]
19. Williams MS. Misdiagnosis: burnout, moral injury, and implications for the electronic health record. J Am Med Inform Assoc 2021;28(5):1047-1050 [FREE Full text] [doi: 10.1093/jamia/ocaa244] [Medline: 33164089]
20. Baxter SL, Saseendrakumar BR, Cheung M, Savides TJ, Longhurst CA, Sinsky CA, et al. Association of electronic health record inbasket message characteristics with physician burnout. JAMA Netw Open 2022;5(11):e2244363 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.44363] [Medline: 36449288]
21. 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 health care interventions: explanation and elaboration. Ann Intern Med 2009;151(4):W65-W94 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00136] [Medline: 19622512]
22. Maslach C, Jackson SE, Leiter MP. Maslach burnout inventory. In: Evaluating Stress. Lanham, MD: Scarecrow Education; 1997.
23. The Stanford Model of Professional Fulfillment™. Stanford Medicine: WellMD & WellPhD. 2016. URL: https://wellmd.stanford.edu/about/model-external.html [accessed 2024-05-01]
24. Shimotsu S, Poplau S, Linzer M. Validation of a brief clinician survey to reduce clinician burnout. J Gen Intern Med 2015;30(2 suppl):S79-S80.
25. Munn Z, Stone JC, Aromataris E, Klugar M, Sears K, Leonardi-Bee J, et al. Assessing the risk of bias of quantitative analytical studies: introducing the vision for critical appraisal within JBI systematic reviews. JBI Evid Synth 2023 Mar 01;21(3):467-471. [doi: 10.11124/JBIES-22-00224] [Medline: 36476419]
26. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25(9):603-605 [FREE Full text] [doi: 10.1007/s10654-010-9491-z] [Medline: 20652370]
27. 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]
28. Higgins JPT, Li T, Deeks JJ. Choosing effect measures and computing estimates of effect. In: Chandler J, Thomas J, Higgins JPT, Page MJ, Cumpston M, Li T, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions, Second Edition. Hoboken: Wiley; 2019:143-176.
29. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
30. Tawfik DS, Phibbs CS, Sexton JB, Kan P, Sharek PJ, Nisbet CC, et al. Factors associated with provider burnout in the NICU. Pediatrics 2017 May;139(5):e20164134 [FREE Full text] [doi: 10.1542/peds.2016-4134] [Medline: 28557756]
31. 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;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
32. Tawfik DS, Sinha A, Bayati M, Adair KC, Shanafelt TD, Sexton JB, et al. Frustration with technology and its relation to emotional exhaustion among health care workers: cross-sectional observational study. J Med Internet Res 2021;23(7):e26817 [FREE Full text] [doi: 10.2196/26817] [Medline: 34255674]

33. Olson K, Sinsky C, Rinne ST, Long T, Vender R, Mukherjee S, et al. Cross-sectional survey of workplace stressors associated with physician burnout measured by the Mini-Z and the maslach burnout inventory. Stress Health 2019;35(2):157-175 [FREE Full text] [doi: 10.1002/smi.2849] [Medline: 30467949]
34. Tai-Seale M, Dillon EC, Yang Y, Nordgren R, Steinberg RL, Nauenberg T, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019;38(7):1073-1078 [FREE Full text] [doi: 10.1377/hlthaff.2018.05509] [Medline: 31260371]
35. Apaydin EA, Rose D, Meredith LS, McClean M, Dresselhaus T, Stockdale S. Association between difficulty with VA patient-centered medical home model components and provider emotional exhaustion and intent to remain in practice. J Gen Intern Med 2020;35(7):2069-2075 [FREE Full text] [doi: 10.1007/s11606-020-05780-8] [Medline: 32291716]
36. Livaudais M, Deng D, Frederick T, Grey-Theriot F, Kroth PJ. Perceived value of the electronic health record and its association with physician burnout. Appl Clin Inform 2022;13(4):778-784 [FREE Full text] [doi: 10.1055/s-0042-1755372] [Medline: 35981548]
37. Tran B, Lenhart A, Ross R, Dorr DA. Burnout and EHR use among academic primary care physicians with varied clinical workloads. AMIA Jt Summits Transl Sci Proc 2019;2019:136-144 [FREE Full text] [Medline: 31258965]
38. Marckini DN, Samuel BP, Parker JL, Cook SC. Electronic health record associated stress: a survey study of adult congenital heart disease specialists. Congenit Heart Dis 2019;14(3):356-361 [FREE Full text] [doi: 10.1111/chd.12745] [Medline: 30825270]
39. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
40. Hilliard RW, Haskell J, Gardner RL. Are specific elements of electronic health record use associated with clinician burnout more than others? J Am Med Inform Assoc 2020;27(9):1401-1410 [FREE Full text] [doi: 10.1093/jamia/ocaa092] [Medline: 32719859]
41. Higgins MCSS, Siddiqui AA, Kosowsky T, Unan L, Mete M, Rowe S, et al. Burnout, professional fulfillment, intention to leave, and sleep-related impairment among radiology trainees across the United States (US): a multisite epidemiologic study. Acad Radiol 2022;29(Suppl 5):S118-S125. [doi: 10.1016/j.acra.2022.01.022] [Medline: 35241358]
42. Czernik Z, Yu A, Pell J, Feinbloom D, Jones CD. Hospitalist perceptions of electronic health records: a multi-site survey. J Gen Intern Med 2022;37(1):269-271 [FREE Full text] [doi: 10.1007/s11606-020-06558-8] [Medline: 33479933]
43. Hauer A, Waukau HJ, Welch P. Physician burnout in Wisconsin: an alarming trend affecting physician wellness. WMJ 2018;117(5):194-200 [FREE Full text] [Medline: 30674095]
44. Gajra A, Bapat B, Jeune-Smith Y, Nabhan C, Klink AJ, Liassou D, et al. Frequency and causes of burnout in US community oncologists in the era of electronic health records. JCO Oncol Pract 2020;16(4):e357-e365 [FREE Full text] [doi: 10.1200/JOP.19.00542] [Medline: 32275848]
45. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
46. Somerson JS, Patton A, Ahmed AA, Ramey S, Holliday EB. Burnout among United States orthopaedic surgery residents. J Surg Educ 2020;77(4):961-968. [doi: 10.1016/j.jsurg.2020.02.019] [Medline: 32171748]
47. Melnick ER, Dyrbye LN, Sinsky CA, Trockel M, West CP, Nedelec L, et al. The association between perceived electronic health record usability and professional burnout among US physicians. Mayo Clin Proc 2020;95(3):476-487 [FREE Full text] [doi: 10.1016/j.mayocp.2019.09.024] [Medline: 31735343]
48. Coleman DM, Money SR, Meltzer AJ, Wohlauer M, Drudi LM, Freischlag JA, et al. Vascular surgeon wellness and burnout: a report from the Society for Vascular Surgery Wellness Task Force. J Vasc Surg 2021;73(6):1841-1850.e3 [FREE Full text] [doi: 10.1016/j.jvs.2020.10.065] [Medline: 33248123]
49. Abraham CM, Zheng K, Norful AA, Ghaffari A, Liu J, Topaz M, et al. Use of multifunctional electronic health records and burnout among primary care nurse practitioners. J Am Assoc Nurse Pract 2021;33(12):1182-1189 [FREE Full text] [doi: 10.1097/JXX.0000000000000533] [Medline: 33534286]
50. Kondrich JE, Han R, Clark S, Platt SL. Burnout in pediatric emergency medicine physicians: a predictive model. Pediatr Emerg Care 2022;38(2):e1003-e1008. [doi: 10.1097/PEC.0000000000002425] [Medline: 35100790]
51. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
52. Mandeville B, Cooper E, Haskell J, Viner-Brown S, Gardner RL. Use of health information technology by Rhode Island physicians and advanced practice providers, 2019. R I Med J (2013) 2020;103(1):21-24 [FREE Full text] [Medline: 32013299]
53. Tiwari V, Kavanaugh A, Martin G, Bergman M. High burden of burnout on rheumatology practitioners. J Rheumatol 2020;47(12):1831-1834 [FREE Full text] [doi: 10.3899/jrheum.191110] [Medline: 32238512]

54. Sinha A, Shanafelt TD, Trockel M, Wang H, Sharp C. Novel nonproprietary measures of ambulatory electronic health record use associated with physician work exhaustion. Appl Clin Inform 2021;12(3):637-646 [FREE Full text] [doi: 10.1055/s-0041-1731678] [Medline: 34261173]
55. Anderson JC, Bilal M, Burke CA, Gaidos JK, Lopez R, Oxentenko AS, et al. Burnout among US gastroenterologists and fellows in training: identifying contributing factors and offering solutions. J Clin Gastroenterol 2023;57(10):1063-1069 [FREE Full text] [doi: 10.1097/MCG.0000000000001781] [Medline: 36477385]
56. Nair D, Brereton L, Hoge C, Plantinga LC, Agrawal V, Soman SS, et al. Burnout among nephrologists in the United States: a survey study. Kidney Med 2022;4(3):100407 [FREE Full text] [doi: 10.1016/j.xkme.2022.100407] [Medline: 35386610]
57. Jha SS, Shah S, Calderon MD, Soin A, Manchikanti L. The effect of COVID-19 on interventional pain management practices: a physician burnout survey. Pain Physician 2020;23(4S):S271-S282 [FREE Full text] [Medline: 32942787]
58. Esmaeilzadeh P, Mirzaei T. Using electronic health records to mitigate workplace burnout among clinicians during the COVID-19 pandemic: field study in Iran. JMIR Med Inform 2021;9(6):e28497 [FREE Full text] [doi: 10.2196/28497] [Medline: 34033578]
59. Holzer KJ, Lou SS, Goss CW, Strickland J, Evanoff BA, Duncan JG, et al. Impact of changes in EHR use during COVID-19 on physician trainee mental health. Appl Clin Inform 2021;12(3):507-517 [FREE Full text] [doi: 10.1055/s-0041-1731000] [Medline: 34077972]
60. Wilkie T, Tajirian T, Thakur A, Mistry S, Islam F, Stergiopoulos V. Evolution of a physician wellness, engagement and excellence strategy: lessons learnt in a mental health setting. BMJ Lead 2023;7(3):182-188 [FREE Full text] [doi: 10.1136/leader-2022-000595] [Medline: 37200187]
61. Eschenroeder HC, Manzione LC, Adler-Milstein J, Bice C, Cash R, Duda C, et al. Associations of physician burnout with organizational electronic health record support and after-hours charting. J Am Med Inform Assoc 2021 Apr 23;28(5):960-966 [FREE Full text] [doi: 10.1093/jamia/ocab053] [Medline: 33880534]
62. Sharp M, Burkart KM, Adelman MH, Ashton RW, Biddison LD, Bosslet GT, et al. A national survey of burnout and depression among fellows training in pulmonary and critical care medicine: a special report by the association of pulmonary and critical care medicine program directors. Chest 2021;159(2):733-742 [FREE Full text] [doi: 10.1016/j.chest.2020.08.2117] [Medline: 32956717]
63. Peccoralo LA, Kaplan CA, Pietrzak RH, Charney DS, Ripp JA. The impact of time spent on the electronic health record after work and of clerical work on burnout among clinical faculty. J Am Med Inform Assoc 2021;28(5):938-947 [FREE Full text] [doi: 10.1093/jamia/ocaa349] [Medline: 33550392]
64. Harris DA, Haskell J, Cooper E, Crouse N, Gardner R. Estimating the association between burnout and electronic health record-related stress among advanced practice registered nurses. Appl Nurs Res 2018;43:36-41. [doi: 10.1016/j.apnr.2018.06.014] [Medline: 30220361]
65. Robertson SL, Robinson MD, Reid A. Electronic health record effects on work-life balance and burnout within the I population collaborative. J Grad Med Educ 2017;9(4):479-484 [FREE Full text] [doi: 10.4300/JGME-D-16-00123.1] [Medline: 28824762]
66. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012;172(18):1377-1385 [FREE Full text] [doi: 10.1001/archinternmed.2012.3199] [Medline: 22911330]
67. Shultz CG, Holmstrom HL. The use of medical scribes in health care settings: a systematic review and future directions. J Am Board Fam Med 2015;28(3):371-381 [FREE Full text] [doi: 10.3122/jabfm.2015.03.140224] [Medline: 25957370]
68. Green-McKenzie J, Somasundaram P, Lawler T, O'Hara E, Shofer FS. Prevalence of burnout in occupational and environmental medicine physicians in the United States. J Occup Environ Med 2020;62(9):680-685 [FREE Full text] [doi: 10.1097/JOM.0000000000001913] [Medline: 32890204]
69. Nguyen OT, Turner K, Charles D, Sprow O, Perkins R, Hong YR, et al. Implementing digital scribes to reduce electronic health record documentation burden among cancer care clinicians: a mixed-methods pilot study. JCO Clin Cancer Inform 2023;7:e2200166 [FREE Full text] [doi: 10.1200/CCI.22.00166] [Medline: 36972488]
70. Ghatnekar S, Faletsky A, Nambudiri VE. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. Health Technol (Berl) 2021;11(4):803-809 [FREE Full text] [doi: 10.1007/s12553-021-00568-0] [Medline: 34094806]
71. Lourie EM, Utidjian LH, Ricci MF, Webster L, Young C, Grenfell SM. Reducing electronic health record-related burnout in providers through a personalized efficiency improvement program. J Am Med Inform Assoc 2021;28(5):931-937 [FREE Full text] [doi: 10.1093/jamia/ocaa248] [Medline: 33166384]
72. DeWitt D, Harrison LE. The potential impact of scribes on medical school applicants and medical students with the new clinical documentation guidelines. J Gen Intern Med 2018;33(11):2002-2004 [FREE Full text] [doi: 10.1007/s11606-018-4582-8] [Medline: 30066114]
73. Liu S, McCoy AB, Wright AP, Carew B, Genkins JZ, Huang SS, et al. Leveraging large language models for generating responses to patient messages—a subjective analysis. J Am Med Inform Assoc 2024:ocae052 [FREE Full text] [doi: 10.1093/jamia/ocae052] [Medline: 38497958]

74. Liu S, McCoy AB, Wright AP, Nelson SS, Huang SS, Ahmad HB, et al. Why do users override alerts? Utilizing large language model to summarize comments and optimize clinical decision support. J Am Med Inform Assoc 2024:ocae041 [FREE Full text] [doi: 10.1093/jamia/ocae041] [Medline: 38452289]
75. Payne TH, Alonso WD, Markiel JA, Lybarger K, Lordon R, Yetisgen M, et al. Using voice to create inpatient progress notes: effects on note timeliness, quality, and physician satisfaction. JAMIA Open 2018;1(2):218-226 [FREE Full text] [doi: 10.1093/jamiaopen/ooy036] [Medline: 31984334]
76. Wang JX, Sullivan DK, Wells AC, Chen JH. ClinicNet: machine learning for personalized clinical order set recommendations. JAMIA Open 2020;3(2):216-224 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa021] [Medline: 32734162]
77. 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]
78. Truhn D, Loeffler CM, Müller-Franzes G, Nebelung S, Hewitt KJ, Brandner S, et al. Extracting structured information from unstructured histopathology reports using Generative Pre-Trained Transformer 4 (GPT-4). J Pathol 2024;262(3):310-319 [FREE Full text] [doi: 10.1002/path.6232] [Medline: 38098169]
79. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
80. Liu S, McCoy AB, Peterson JF, Lasko TA, Sittig DF, Nelson SD, et al. Leveraging explainable artificial intelligence to optimize clinical decision support. J Am Med Inform Assoc 2024;31(4):968-974 [FREE Full text] [doi: 10.1093/jamia/ocae019] [Medline: 38383050]
81. Atutxa A, Perez A, Casillas A, Atutxa A, Perez A, Casillas A. Machine learning approaches on diagnostic term encoding with the ICD for clinical documentation. IEEE J Biomed Health Inform 2018;22(4):1323-1329 [FREE Full text] [doi: 10.1109/JBHI.2017.2743824] [Medline: 28858819]
82. Khairat S, Coleman C, Ottmar P, Jayachander DI, Bice T, Carson SS. Association of electronic health record use with physician fatigue and efficiency. JAMA Netw Open 2020;3(6):e207385 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.7385] [Medline: 32515799]
83. Murphy DR, Satterly T, Giardina TD, Sittig DF, Singh H. Practicing clinicians' recommendations to reduce burden from the electronic health record inbox: a mixed-methods study. J Gen Intern Med 2019;34(9):1825-1832 [FREE Full text] [doi: 10.1007/s11606-019-05112-5] [Medline: 31292905]
84. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res 2023;25:e48009 [FREE Full text] [doi: 10.2196/48009] [Medline: 37566454]
85. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30(7):1237-1245 [FREE Full text] [doi: 10.1093/jamia/ocad072] [Medline: 37087108]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
EMR: electronic medical record
IV: inverse variation methods
JBI: Joanna Briggs Institute
LLM: large language model
MBI-HSS: Maslach Burnout Inventory-Human Services Survey instrument
NLP: natural language processing
NOS: Newcastle-Ottawa Scale
OR: odds ratio
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis

Edited by C Lovis; submitted 27.11.23; peer-reviewed by JY Nam, J Wolff, I Mircheva, R Koppel; comments to author 14.01.24; revised version received 23.02.24; accepted 17.04.24; published 12.06.24.

Please cite as:
Wu Y, Wu M, Wang C, Lin J, Liu J, Liu S
Evaluating the Prevalence of Burnout Among Health Care Professionals Related to Electronic Health Record Use: Systematic Review and Meta-Analysis
JMIR Med Inform 2024;12:e54811
URL: https://medinform.jmir.org/2024/1/e54811
doi:10.2196/54811
PMID:38865188

How to Elucidate Consent-Free Research Use of Medical Data: A Case for "Health Data Literacy"

Gesine Richter[1,2], PhD; Michael Krawczak[3], MSc, PhD

1

2

3

Corresponding Author:
Gesine Richter, PhD

Abstract

The extensive utilization of personal health data is one of the key success factors of modern medical research. Obtaining consent to the use of such data during clinical care, however, bears the risk of low and unequal approval rates and risk of consequent methodological problems in the scientific use of the data. In view of these shortcomings, and of the proven willingness of people to contribute to medical research by sharing personal health data, the paradigm of informed consent needs to be reconsidered. The European General Data Protection Regulation gives the European member states considerable leeway with regard to permitting the research use of health data without consent. Following this approach would however require alternative offers of information that compensate for the lack of direct communication with experts during medical care. We therefore introduce the concept of "health data literacy," defined as the capacity to find, understand, and evaluate information about the risks and benefits of the research use of personal health data and to act accordingly. Specifically, health data literacy includes basic knowledge about the goals and methods of data-rich medical research and about the possibilities and limits of data protection. Although the responsibility for developing the necessary resources lies primarily with those directly involved in data-rich medical research, improving health data literacy should ultimately be of concern to everyone interested in the success of this type of research.

(JMIR Med Inform 2024;12:e51350) doi:10.2196/51350

KEYWORDS

health data literacy; informed consent; broad consent; data sharing; data collection; data donation; data linkage; personal health data

Data-Rich Research, Broad Consent, and Informedness

Various initiatives around the world are currently working on the technical and organizational requirements to make data from different sources and contexts usable for medical research (eg, MyHealthRecord in Australia, FINDATA in Finland, and the Medical Informatics Initiative in Germany). The starting points of these endeavors often are local, regional, or national health care data repositories that must nevertheless be highly linkable to allow full exploitation of their scientific value. This connectivity requirement implies that the data cannot be fully anonymized before being moved into the research domain.

One of the ethical prerequisites for research on humans—and thus for research using identifiable personal health data—is the informed consent of the data subjects. However, being properly informed requires that those affected (1) are capable of making self-determined decisions in the first place; (2) were informed about the nature, benefits, and risks of the research in question; (3) have understood the importance of this information; and (4) are able to decide voluntarily and without coercion for or against participation.

Not least because of the increasing relevance of hypotheses-free research approaches (keyword: big data), the storage and use of data for future, currently undeterminable purposes also play an increasingly important role in medical research. Recent studies have shown that patients and members of the general public are very willing to share personal health data for research (eg, [1]), even if no information about the purposes and aims of the research can be provided at the time consent is given. Notably, this attitude turned out to be mainly motivated by altruism, solidarity, and the idea of reciprocity. Since the paradigm of project-related informed consent is difficult to transfer to such unspecific practice, the World Medical Association changed its regulations on research with identifiable data when revising the Declaration of Helsinki in 2013 [2]. There was no longer a requirement for specific information about the subjects of future research, thereby paving the way for a new form of "broad consent."

In essence, "broad consent" means the one-off, unspecific agreement to the use of one's personal data for medical research without knowing who will access the data when and to what end. However, since the data in question are usually collected in a clinical care context, the suitability and practicality of broad

consent as a legitimation for their research use is limited. First, the temporal and spatial linking of the consent process to care measures can lead to incorrect therapeutic [3] and diagnostic [4] assumptions on the side of the patient. Second, in the time available, it is hardly possible to create sufficient understanding of the benefits and risks of the envisaged research, despite great efforts to ensure that the corresponding information and consent documents are legible. Finally, asking for consent during clinical care bears a substantial risk of low and unequal approval rates, which can lead to methodological problems in the scientific use of the data.

In view of these shortcomings, and of the proven willingness of people to contribute to medical research by sharing personal health data, the means to achieve practically feasible and truly informed consent needs to be reconsidered. In particular, is consent-free data use for medical research, combined with the possibility of straightforward opt-out by the data subjects after thorough consideration, a better option for legitimizing the secondary use of health data? This question is all the more justified as numerous studies in the United Kingdom, Iceland, Norway, Sweden, and Germany, among others, have shown a generally positive attitude of people toward such a regulation (eg, United Kingdom [5]; United Kingdom, Iceland, Norway, and Sweden [1]; Norway [6]; and Germany [7,8]).

In the following, we will first introduce "data donation" as an opt-out approach to legitimizing the secondary research use of personal medical data. Since opt-out would imply that patients are no longer informed directly about the research-associated risks and benefits, alternative ways of information provision must be explored in the context of data donation if the paradigm of informedness was to be maintained. We therefore also introduce the concept of "health data literacy," defined as the capacity to find, understand, and evaluate information about data-rich medical research. Although a case for general health data literacy can be made independently of the issue of patient consent, its consideration becomes particularly urgent for the latter if the framework of consenting was to change from opt-in to opt-out.

Data Donation: Consent-Free Research Use of Medical Data Plus Opt-Out

The European General Data Protection Regulation (EU-GDPR) gives European member states considerable leeway with regard to permitting the research use of health data without consent. While Article 9 Paragraph 1 of the EU-GDPR clearly prohibits the processing of personal genetic, biometric, or health data, Article 9 Paragraph 2(j) explicitly exempts processing for scientific research purposes [9]. In addition, Article 89 allows national legislation to provide for this exception, subject to appropriate safeguards for the rights and freedom of the data subjects.

In Germany, the ethical, legal, technological, and organizational framework of the consent-free use of health data was examined in 2020 in a detailed report to the Federal Ministry of Health [10]. In addition to its legal admissibility, the report addressed the scientific benefits of such an approach, its impact upon the right of informational self-determination, and the necessity and possibilities for fair involvement of the data subjects. The authors concluded that it would be possible in Germany to replace the requirement for explicit consent for research with personal medical data by an equivalent legal permission, combined with an easy-to-exercise opt-out. Under certain conditions, such "data donation" (as it was termed in the report) would be both legally possible and ethically reasonable.

The above notwithstanding, the authors were also unequivocal that the actual process of data access by potential users should be independent of whether access is legitimized by opt-in or opt-out. The involvement of an ethics board or a use-and-access committee that reviews and decides data applications remains essential in both cases. Notably, such institutions also play an important role in weighing the potential risks and benefits of individual research projects, a legitimation mechanism that was deliberately placed on the same level as consent by the EU-GDPR.

Importantly with a view to the following considerations, the report clarified that, in addition to technical and organizational protective measures, one prerequisite for the acceptability of data donation would be that patients and citizens were sufficiently well informed about it. This proviso inevitably leads to the question of how sufficient knowledgeability can be achieved if the decision about sharing one's data for research purposes is no longer made actively, following thorough verbal explanation, but passively by exercising or not exercising a right of objection.

Limits of Top-Down "Informability": the COVID-19 Infodemic as an Example

Since data donation, in the above sense, would be temporally and spatially decoupled from medical care and instead be anchored in everyday life, alternative offers of information would have to compensate for the lack of direct communication with medical or scientific experts [11]. Yet, the COVID-19 pandemic recently highlighted that the expansion of top-down media campaigns alone is not sufficient to adequately convey the complex aspects of medical research to the general public. Instead, it turned out that, despite the general increase in information provided, many people who opposed vaccination in the first place still were not sufficiently receptive to scientific facts [12]. Moreover, even some kind of social grouping occurred along people's vaccination status, and the COSMO study carried out in Germany and Austria revealed that the stronger the identification with being unvaccinated, the lower the inclination to change this status, and the greater the feeling of discrimination [13]. Obviously, the ability to become informed ("informability") had reached its limits in view of the amount of information available, a paradox that lamentably also had a negative impact upon the effectiveness of public health measures taken.

In connection with the COVID-19 pandemic, the World Health Organization (WHO) coined the term "infodemic" for the increasingly observed susceptibility of people to fake news as a result of reduced informability. According to the WHO, the

infodemic caused a high degree of uncertainty in the population, a greater willingness to engage in health-damaging and risk-taking behavior, and an increased distrust of the health authorities [14]. The "Infodemic Management" called for by the WHO aimed to enable the population to better understand information from health experts and to become more resistant to misinformation [15].

Ways to Better Informability?

In view of its complexity, it seems unrealistic to convey all relevant information about the research use of personal health data at once. We therefore propose "health data literacy" as a basis for better informability of the general population and, hence, as a means to uphold the paradigm of informed consent even in the context of data donation in the above sense. For a well-informed general public, data donation would indeed mean nothing more than a change in decision format—from opt-in to opt-out.

In a narrower sense, the word "literacy" stands for the ability to read and, thereby, to acquire education and knowledge. According to the Organisation for Economic Co-operation and Development (OECD), understanding and interpreting written material should enable citizens to develop their own potential and to fully participate in societal affairs [16]. The starting point of our considerations on health data literacy therefore will be a class of communication models that focus upon the possible causes of limited informability.

One decisive factor for the success of communication is the thought system of the recipient. Since we often have little time to consider large amounts of everyday information, we believe statements that we have heard very often to be more credible than others [17]. This effect is reinforced by the phenomenon of group polarization: those who share a widespread opinion on complex issues are more likely to be reserved about new information and tend to believe whatever confirms their own viewpoint rather than information that does not fit. This selective form of information intake can, for example, increase polarization in social disputes even in the presence of reliable evidence and information [18]. The concept of health data literacy picks up on the basic idea of these communication models and aims to create anchor points in the knowledge base of people, where information on the benefits and risks of data-rich medical research can be stored and evaluated.

Value congruence approaches aim in a similar direction, in that they try to increase trust in certain institutions [19,20]. Such trust will be greater when more individuals perceive that their interests and values are shared by the institution in question, because trust is also largely based upon the perception of common values. This applies all the more to institutions that use health data for research, and it is therefore in the best interest of such institutions to develop and represent values that are highly rated by the public [19,20]. In this context, widespread health data literacy could form the breeding ground for the perception of a congruence of values and, thus, for greater trust in the recipients and beneficiaries of data donation.

The Concept of "Health Data Literacy"

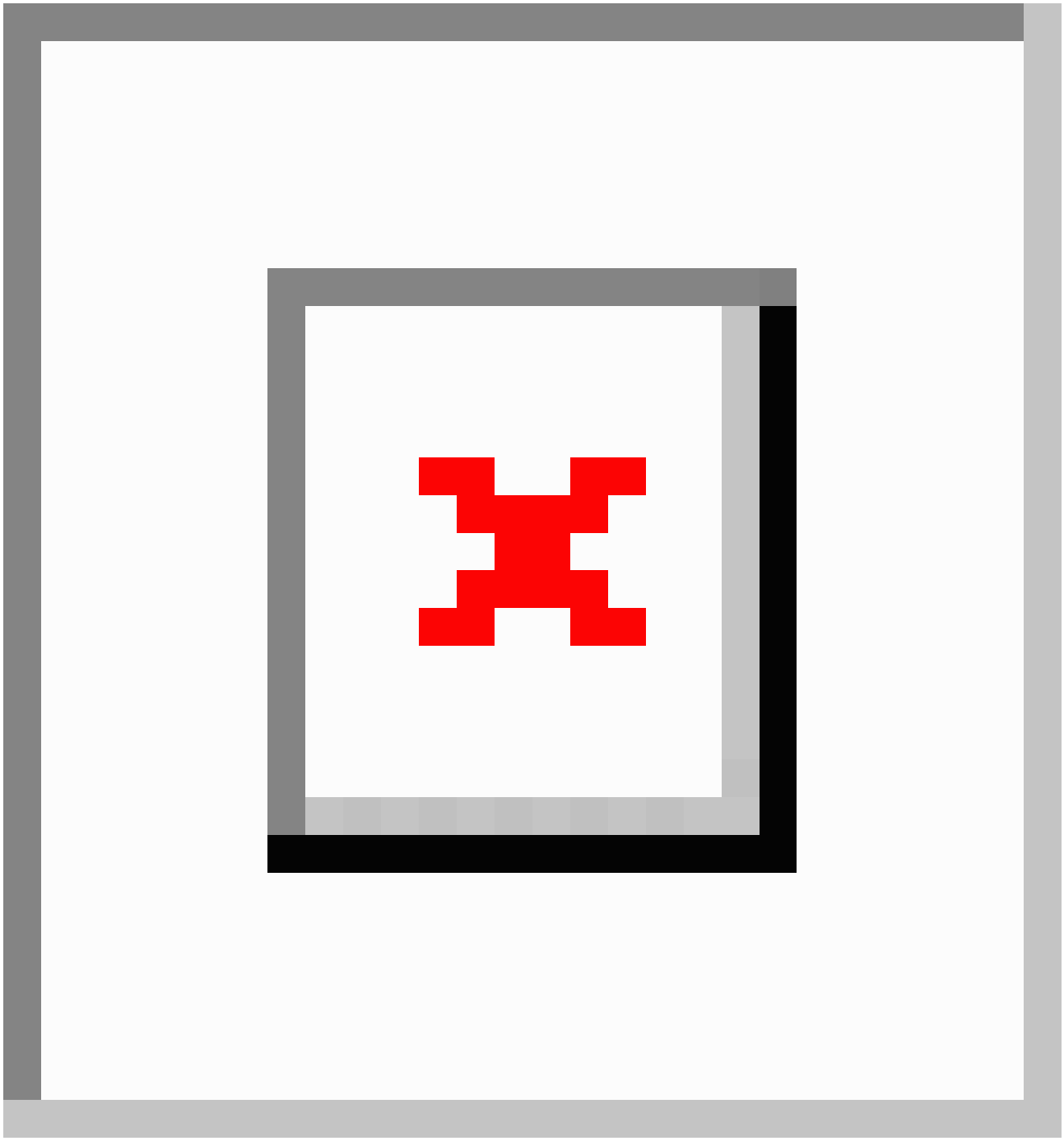
An individual's health data literacy is positioned between their health literacy and their data literacy, where the latter in particular has been promoted politically, for example, by the data strategy of the German federal government [21].

- In view of the increasingly specific treatment options promised by so-called "precision medicine," citizens would be well advised to take an interest in issues related to disease prevention and medical care [22]. The associated term "health literacy" summarizes both the motivation and the ability to find, understand, evaluate, and apply the information underlying personal health–related decisions [23]. Numerous international studies have measured and compared the level of health literacy in different populations (eg, [24]), as well as spurring considerations as to how health literacy can be increased (eg, [25]).
- The term "data literacy" refers to knowledge about data and their use in general, including legal, ethical, and social aspects. Data literacy thus forms the basis of personal self-determination in an increasingly digitalized society [26]. The aim of data literacy is an ability to weigh one's own personal rights against the potential benefits of making personal data available to others [27].

In combining both abovementioned terms, "health data literacy" stands for the capacity to find, understand, and evaluate information about the risks and benefits of medical research with personal health data; to compare this information with one's own values; and to act accordingly. Health data literacy is thus a transformer of information into informed action, aimed at a level of thematic familiarity that enables self-determined decision-making about the sharing of one's own health data with the research community. Specifically, health data literacy should at the very least include basic knowledge about the goals and methods of data-rich medical research and about the possibilities and limits of data protection.

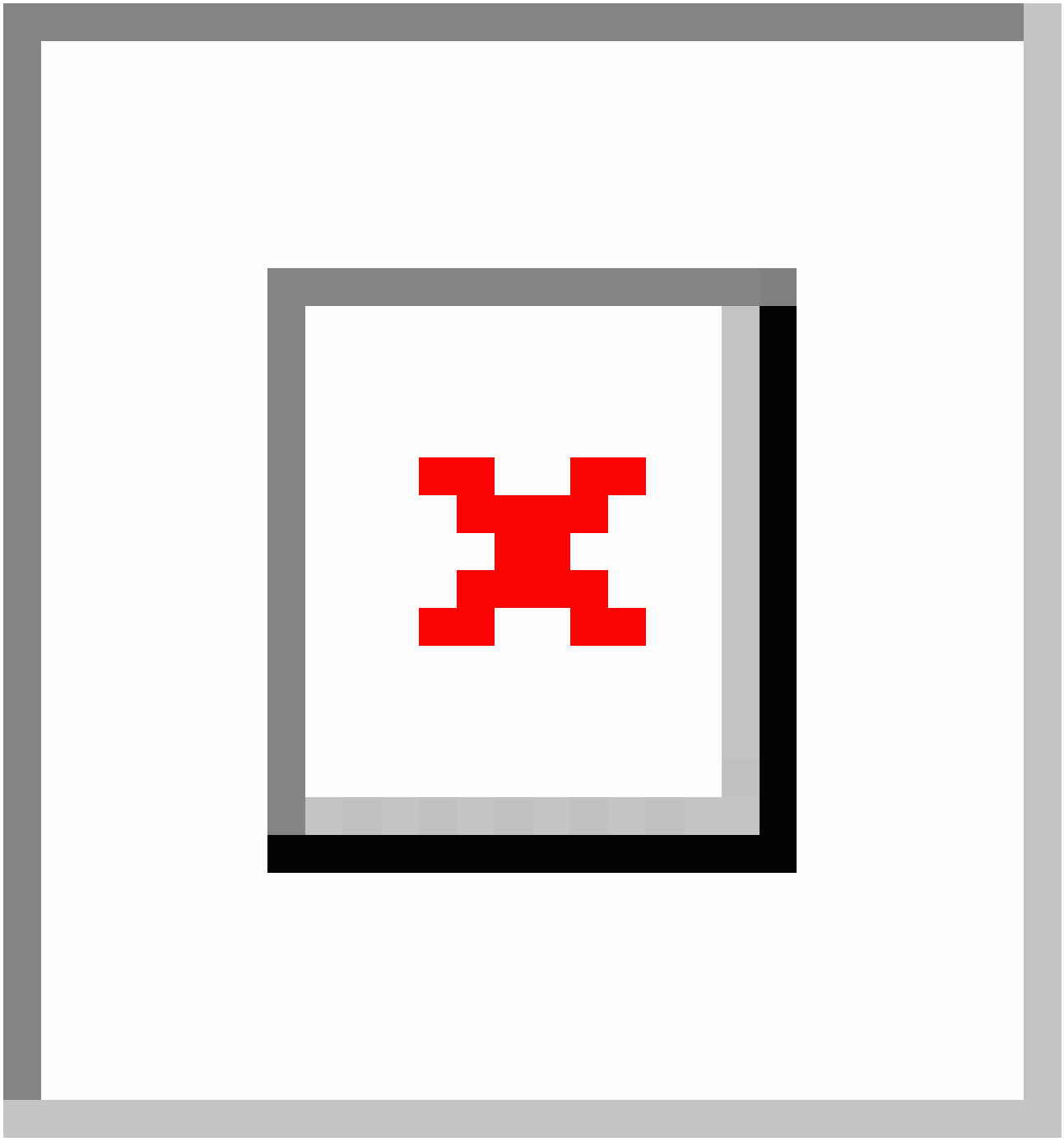
The increasing relevance of personal health data for medical research has led to a large number of measures to increase the societal acceptance of the use of such data. However, legislative regulations on data governance and data protection, as well as efforts to increase patient involvement and public information, are likely to have greater impact when they are met with more adequate prior knowledge in the sense of health data literacy (Figure 1).

Figure 1. Health data literacy as a breeding ground for the societal acceptance of data donation.



When the mental anchor points set by health data literacy receive information on scientific successes, new technical and organizational developments, as well as possible setbacks of data-rich medical research (keyword: transparency), this information can be evaluated competently by the recipient and compared to their own expectations. In the aftermath of such reflections, informed self-determination and sufficient trust in regulations and institutions can develop (Figure 2).

Figure 2. Transparency nourishes confidence and trust in data-rich medical research.



Outlook: Feasibility and Implementation of Health Data Literacy

Numerous international studies among patients and in the general population have revealed a broad positive attitude toward the provision of personal health data to medical research (eg, [28]). This approval was consistently found to be driven by a sense of reciprocity, that is, a wish to give something back after benefiting from research (eg, [29,30]). Evidence also emerged for the widespread belief in a social duty of citizens to contribute their own data to research, independent of their personal benefit [31,32]. At the same time, however, a craving for more detailed information was observed, up to and including the view that every individual is responsible themselves to find out about the nature and benefits of research with personal health data (eg, [33]).

In summary, we are thus in a situation where (1) there is little doubt about the need to utilize personal health data from different contexts to achieve the goals of modern medical research, (2) the consent-free use of such data meets broad approval by the general public, and (3) there is a widespread willingness of people to acquire the knowledge necessary to make a self-determined decision about data donation. The most compelling argument for general health data literacy is therefore self-evident: widespread background knowledge of the risks and benefits of data-rich medical research would allow the paradigm of informedness to be maintained even if consent to

participation in research is implement by opt-out, rather than opt-in.

However, the appeal of general health data literacy undoubtedly goes beyond the issue of data donation. Its necessity arises from the increasing complexity of data-rich medical research, which can no longer be explained adequately via waiting room leaflets or doctor consultations. We are also aware that improved health data literacy could, in principle, help to reduce some of the misunderstandings of patients that we somehow held against broad consent when advocating data donation. However, in view of the many advantages of data donation summarized above, we think that only little importance should be attached to this possibility.

Attempts to establish general health data literacy should strive for a certain level of competence across as broad a proportion of the population as possible. This goal not only expresses fairness and ensures equal representation of different societal groups in medical research but can also help to reduce the vulnerability to fake information as a potential threat to public health, as observed during the COVID-19 pandemic. Achieving equity in practice will require the development and provision of target group–specific offers of information and education. One particularly efficient way to increase health data literacy across the board would be to start this process in school, as suggested previously to strengthen health literacy [25]. This approach is not only easy to implement in practice; it would also offer the opportunity to use children as multipliers among friends and family.

Further research is needed to determine exactly what kind of information should be communicated, in what form, and to whom to improve health data literacy in a given population. These questions are ideally answered through cocreation research involving representatives of different target groups to enhance the credibility of the education curriculum and content among end users. However, although the responsibility for developing the necessary resources lies primarily with those directly involved in data-rich medical research, improving health data literacy should ultimately be of concern to everyone interested in the success of this type of research.

Acknowledgments

We acknowledge financial support by Deutsche Forschungsgemeinschaft (DFG) within the funding program Open Access-Publikationskosten. We are most grateful to Claudia Bozzaro, Kiel University, for discussing health data literacy with us.

Authors' Contributions

The idea of health data literacy was first conceived by GR; GR and MK jointly developed the concept further and authored the manuscript.

Conflicts of Interest

None declared.

References

1. Viberg Johansson J, Bentzen HB, Shah N, et al. Preferences of the public for sharing health data: discrete choice experiment. JMIR Med Inform 2021 Jul 5;9(7):e29614. [doi: 10.2196/29614] [Medline: 36260402]
2. WMA Declaration of Helsinki – ethical principles for medical research involving human subjects. World Medical Association. 2022 Sep 6. URL: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ [accessed 2024-01-18]
3. Appelbaum PS, Roth LH, Lidz CW, Benson P, Winslade W. False hopes and best data: consent to research and the therapeutic misconception. Hastings Cent Rep 1987 Apr;17(2):20-24. [Medline: 3294743]
4. Nobile H, Vermeulen E, Thys K, Bergmann MM, Borry P. Why do participants enroll in population biobank studies? a systematic literature review. Expert Rev Mol Diagn 2013 Jan;13(1):35-47. [doi: 10.1586/erm.12.116] [Medline: 23256702]
5. Jones LA, Nelder JR, Fryer JM, et al. Public opinion on sharing data from health services for clinical and research purposes without explicit consent: an anonymous online survey in the UK. BMJ Open 2022 Apr 27;12(4):e057579. [doi: 10.1136/bmjopen-2021-057579] [Medline: 35477868]
6. Eikemo H, Roten LT, Vaaler AE. Research based on existing clinical data and biospecimens: a systematic study of patients' opinions. BMC Med Ethics 2022 Jun 16;23(1):60. [doi: 10.1186/s12910-022-00799-4] [Medline: 35710552]
7. Richter G, Trigui N, Caliebe A, Krawczak M. Attitude towards consent-free research use of personal medical data in the general German population. Heliyon 2024 Mar 11;10(6):e27933. [doi: 10.1016/j.heliyon.2024.e27933] [Medline: 38509969]
8. Köngeter A, Schickhardt C, Jungkunz M, Bergbold S, Mehlis K, Winkler EC. Patients' willingness to provide their clinical data for research purposes and acceptance of different consent models: findings from a representative survey of patients with cancer. J Med Internet Res 2022 Aug 25;24(8):e37665. [doi: 10.2196/37665] [Medline: 36006690]
9. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive

95/46/EC (General Data Protection Regulation). EUR-Lex. URL: http://data.europa.eu/eli/reg/2016/679/2016-05-04 [accessed 2024-01-18]
10. Strech D, von Kielmansegg S, Zenker S, Krawczak M, Semler SC. Wissenschaftliches Gutachten „Datenspende“ – Bedarf für die Forschung, ethische Bewertung, rechtliche, informationstechnologische und organisatorische Rahmenbedingungen [Article in German]. Bundesministerium für Gesundheit. 2020 Mar 30. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Ministerium/Berichte/Gutachten_Datenspende.pdf [accessed 2024-01-18]
11. Jungkunz M, Köngeter A, Spitz M, et al. Stellungnahme zur Etablierung der sekundären Forschungsnutzung von Behandlungsdaten in Deutschland: Ergebnisse des Verbundprojekts LinCDat: „Learning from Clinical Data. Ethical, Social and Legal Aspects" [Article in German]. Forum Marsilius-Kolleg 2022 Nov 24;21. [doi: 10.11588/fmk.2022.1.91697]
12. Rathore FA, Farooq F. Information overload and infodemic in the COVID-19 pandemic. J Pak Med Assoc 2020 May;70(Suppl 3)(5):S162-S165. [doi: 10.5455/JPMA.38] [Medline: 32515403]
13. COSMO PANEL—Langzeitstudie zum Erleben und Verhalten von Geimpften und Ungeimpften in Deutschland und Österreich [Article in German]. COSMO. 2022 Mar 18. URL: https://projekte.uni-erfurt.de/cosmo2020/web/summary/panel2/ [accessed 2022-12-13]
14. Infodemic. World Health Organization. URL: https://www.who.int/health-topics/infodemic#tab=tab_1 [accessed 2023-07-21]
15. 1st WHO infodemic manager training. World Health Organization. 2020 Nov. URL: https://www.who.int/teams/epi-win/infodemic-management/1st-who-training-in-infodemic-management [accessed 2023-07-21]
16. Adult literacy. Organisation for Economic Co-operation and Development (OECD). URL: https://www.oecd.org/education/innovation-education/adultliteracy.htm [accessed 2023-07-21]
17. Kahneman D. Schnelles Denken, Langsames Denken: Siedler Verlag, München; 2021.
18. Lord CG, Ross L, Lepper MR. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. J Pers Soc Psychol 1979 Nov;37(11):2098-2109. [doi: 10.1037//0022-3514.37.11.2098]
19. Sheehan M, Friesen P, Balmer A, et al. Trust, trustworthiness and sharing patient data for research. J Med Ethics 2020 May 18:medethics-2019-106048. [doi: 10.1136/medethics-2019-106048] [Medline: 32424061]
20. Holland S, Cawthra J, Schloemer T, Schröder-Bäck P. Trust and the acquisition and use of public health information. Health Care Anal 2022 Mar;30(1):1-17. [doi: 10.1007/s10728-021-00436-y] [Medline: 34751865]
21. Datenstrategie der Bundesregierung: Eine Innovationsstrategie für gesellschaftlichen Fortschritt und nachhaltiges Wachstum - Kabinettfassung [Article in German]. Bundesregierung. 2021 Jan 27. URL: https://www.publikationen-bundesregierung.de/pp-de/publikationssuche/datenstrategie-der-bundesregierung-1845632 [accessed 2023-07-21]
22. Budin-Ljøsne I, Harris JR. Ask not what personalized medicine can do for you--ask what you can do for personalized medicine. Public Health Genomics 2015 Mar 6;18(3):131-138. [doi: 10.1159/000373919] [Medline: 25766382]
23. Sørensen K, van den Broucke S, Fullam J, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health 2012 Jan 25;12:80. [doi: 10.1186/1471-2458-12-80] [Medline: 22276600]
24. Sørensen K, Pelikan JM, Röthlin F, et al. Health literacy in Europe: comparative results of the European Health Literacy Survey (HLS-EU). Eur J Public Health 2015 Dec;25(6):1053-1058. [doi: 10.1093/eurpub/ckv043] [Medline: 25843827]
25. Schaeffer D, Hurrelmann K, Bauer U. Nationaler Aktionsplan Gesundheitskompetenz Die Gesundheitskompetenz in Deutschland Stärken: KomPart Verlagsgesellschaft mbH; 2018.
26. Renz A, Etsiwah B, Burgueno Hopf AT. Datenkompetenz. Whitepaper: Weizenbaum-Institut für die vernetzte Gesellschaft; 2021.
27. Hummel P, Braun M, Augsberg S, von Ulmenstein U, Dabrock P. Datensouveränität Governance-Ansätze Für Den Gesundheitsbereich: Springer; 2021:11. [doi: 10.1007/978-3-658-33755-1]
28. Aitken M, de St Jorre J, Pagliari C, Jepson R, Cunningham-Burley S. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. BMC Med Ethics 2016 Nov 10;17(1):73. [doi: 10.1186/s12910-016-0153-x] [Medline: 27832780]
29. Richter G, Krawczak M, Lieb W, Wolff L, Schreiber S, Buyx A. Broad consent for health care-embedded biobanking: understanding and reasons to donate in a large patient sample. Genet Med 2018 Jan;20(1):76-82. [doi: 10.1038/gim.2017.82] [Medline: 28640237]
30. A public dialogue on genomic medicine: time for a new social contract? Ipsos MORI. 2019. URL: https://www.ipsos.com/sites/default/files/ct/publication/documents/2019-04/public-dialogue-on-genomic-medicine-full-report.pdf [accessed 2024-01-18]
31. Skatova A, Goulding J. Psychology of personal data donation. PLoS One 2019 Nov 20;14(11):e0224240. [doi: 10.1371/journal.pone.0224240] [Medline: 31747408]
32. Richter G, Borzikowsky C, Hoyer BF, Laudes M, Krawczak M. Secondary research use of personal medical data: patient attitudes towards data donation. BMC Med Ethics 2021 Dec 15;22(1):164. [doi: 10.1186/s12910-021-00728-x] [Medline: 34911502]
33. Platt J, Raj M, Büyüktür AG, et al. Willingness to participate in health information networks with diverse data use: evaluating public perspectives. EGEMS (Wash DC) 2019 Jul 25;7(1):33. [doi: 10.5334/egems.288] [Medline: 31367650]

Abbreviations

EU-GDPR: European General Data Protection Regulation
OECD: Organisation for Economic Co-operation and Development
WHO: World Health Organization

Edited by C Lovis; submitted 28.07.23; peer-reviewed by G Arnolda, LD C, S McLennan, S Wiertz; revised version received 19.01.24; accepted 21.04.24; published 18.06.24.

Please cite as:
Richter G, Krawczak M
How to Elucidate Consent-Free Research Use of Medical Data: A Case for "Health Data Literacy"
JMIR Med Inform 2024;12:e51350
URL: https://medinform.jmir.org/2024/1/e51350
doi:10.2196/51350

Viewpoint

Using a Natural Language Processing Approach to Support Rapid Knowledge Acquisition

Taneya Y Koonce[1*], MSLS, MPH; Dario A Giuse[2*], MS, Dr Ing; Annette M Williams[1], MLS; Mallory N Blasingame[1], MA, MSIS; Poppy A Krump[1], MSIS; Jing Su[1], MS, MSIS, MD; Nunzia B Giuse[1,2], MLS, MD

[1]Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN, United States

[2]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

[*]these authors contributed equally

Corresponding Author:
Taneya Y Koonce, MSLS, MPH
Center for Knowledge Management
Vanderbilt University Medical Center
3401 West End
Suite 304
Nashville, TN, 37203
United States
Phone: 1 6159365790
Email: taneya.koonce@vumc.org

Abstract

Implementing artificial intelligence to extract insights from large, real-world clinical data sets can supplement and enhance knowledge management efforts for health sciences research and clinical care. At Vanderbilt University Medical Center (VUMC), the in-house developed Word Cloud natural language processing system extracts coded concepts from patient records in VUMC's electronic health record repository using the Unified Medical Language System terminology. Through this process, the Word Cloud extracts the most prominent concepts found in the clinical documentation of a specific patient or population. The Word Cloud provides added value for clinical care decision-making and research. This viewpoint paper describes a use case for how the VUMC Center for Knowledge Management leverages the condition-disease associations represented by the Word Cloud to aid in the knowledge generation needed to inform the interpretation of phenome-wide association studies.

(JMIR Med Inform 2024;12:e53516) doi:10.2196/53516

KEYWORDS

natural language processing; electronic health records; machine learning; data mining; knowledge management; NLP

Introduction

The rapid advancement and availability of artificial intelligence (AI) approaches provide biomedical informatics groups with opportunities for exploring and generating insights from internal and external data at scale to enhance health sciences research and clinical care [1,2]. One such opportunity is using natural language processing (NLP) to extract usable knowledge from the vast amounts of structured and unstructured clinical data captured daily via the electronic health record (EHR). Insights from this process can be used to inform patient care, target information provision, and generate research hypotheses. This paper presents some of the activities that such usable knowledge makes possible.

Vanderbilt University Medical Center (VUMC) maintains an electronic health repository containing data for over 4.6 million individuals, going back to 1995, which includes structured data (eg, laboratory results and vital signs), textual data (eg, provider notes and radiology interpretations), reports (eg, electrocardiograms and pulmonary function test results), and image data. Included in this vendor-agnostic repository are all VUMC patient data captured from the in-house developed StarPanel EHR (VUMC) dating back to 2001 [3] and VUMC's current vendor-based EHR (Epic; Epic Systems Corporation), which was implemented in 2017 [4]. Roughly 850,000 new documents are added daily.

To identify and quickly represent the most critical information about a particular patient or population from this large data set, VUMC established the Word Cloud, a real-time and at-scale concept extraction tool that uses NLP to create a visual, time-oriented representation of clinical data [5-7]. The Word Cloud NLP uses a rules-based, finite-state machine approach

to process all nonimage incoming documents in real time and extract coded concepts using the Unified Medical Language System (UMLS) terminology [8]. With a processing speed of more than 50,000 documents per minute, the Word Cloud NLP is faster than currently available concept extraction NLP tools such as Apache cTakes (50,000 documents per hour; Apache Software Foundation, Mayo Clinic) [9] and MetaMap (22 citations per minute; National Library of Medicine) [10]. The rapid speed allows for better integration into the clinical workflow as real time–generated Word Cloud concepts are immediately presented to health care providers as they access the feature in the medical chart. The system handles all linguistic phenomena in clinical text, including acronyms, abbreviations, misspellings, negation, family history, uncertainty, and differential diagnosis. Excluding image data, the entire EHR repository is included in the Word Cloud NLP database, which uses close to 14,000 UMLS concepts to index 1.7 billion documents. In addition to the individual patient concepts, which include pointers to the original documents, the database also includes population-level associations of any pair of concepts.

The original purpose of the Word Cloud data was to provide a user interface that displays all concepts extracted from a patient's clinical documents in a word cloud display, with the size of each concept indicating how often the concept was documented for the patient. This interface is available to all users of the EHR. The Word Cloud data have been used since 2019 to generate clinical alerts for a variety of situations, such as flagging patients with implanted cardiac devices and a positive blood culture, patients with signs of serious inflammation due to immune checkpoint inhibitors, or patients with Andersen-Tawil syndrome who might be candidates for enrollment into a research study. The Word Cloud data drive real-time decision support by injecting detected concepts back into the VUMC EHR [11]. Because all the concepts extracted by the Word Cloud NLP are stored in the enterprise data lake, these data are also available for retrospective research and can be easily combined with other data such as the International Statistical Classification of Diseases codes or coded medications data [11].

The Center for Knowledge Management (CKM) has explored how the Word Cloud can be leveraged by information scientists engaged in EHR projects. The CKM facilitates the discovery and integration of external knowledge into medical practice and promotes curation, archiving, and reuse of internal knowledge across VUMC [12-15]. This viewpoint paper details how the CKM's innovative application of the Word Cloud enhances knowledge generation processes and describes future directions for NLP in knowledge management.

Case Description

Collaborations with medical center researchers comprise the majority of the CKM's partnership activities. A recent project to inform the interpretation of phenome-wide association studies (PheWASs) using evidence-linked knowledge bases illustrates these types of partnerships [16]. PheWASs examine relationships between markers (genetic or nongenetic) and phenotypes, producing extensive lists of possibly relevant marker-phenotype associations [17,18]. A methodological approach to compare known associations with PheWAS results can make it easier to identify potentially novel PheWAS outcomes [16]. Knowledge bases—created in part from synthesized evidence sources and primary literature documenting disease causes, risks, and complications—can be used for these comparisons.

For this research collaboration, the CKM created a "condition flowchart" with the causes, risk factors, and complications of a given medical condition. The sources consulted to create the flowchart include evidence synthesis resources (eg, UpToDate; UpToDate, Inc), medical textbooks (eg, Goldman-Cecil Medicine), and consumer health websites (eg, MedlinePlus; National Library of Medicine). From each source, the CKM team identified all causes, risk factors, and complications for the condition of interest and added them to the flowchart. Our collaborators then used the flowchart to create a knowledge base of phecodes for the PheWAS analysis. During flowchart creation, the CKM leveraged the Word Cloud to identify meaningful disease-condition associations—based on real-world population-level data—and target appropriate primary literature to substantiate the observed linkages.

Identifying Meaningful Condition Associations From the EHR

Each flowchart focuses on a specific clinical condition (eg, hypertension and hypotension), which is searched against the Word Cloud. Using a population-level analysis feature, the Word Cloud returns a list of all UMLS concepts represented in the EHR records of patients with the specified condition. The expected value is calculated for each UMLS concept [19] and the ratio of actual-to-expected patient cases (ie, strength of association) is then used to rank the list of causes, risk factors, and complications on the flowchart. This ranking thus provides our team with rapid knowledge acquisition of what is associated with the condition of interest. The actual-to-expected ratio for concept 1 and concept 2 is computed as follows:

where T=total population size, $a_{1,2}$=number of patients with both concept 1 and concept 2, n_1=number of patients with concept 1, and n_2=number of patients with concept 2.

A strength of association ratio of 15 or higher indicates that the concept occurs more often than expected by chance and signifies a meaningful relationship between the UMLS concept and the condition. Figure 1 provides an example of the UMLS concepts most associated with orthostatic hypotension in 71,996 patients.

Figure 1. Snapshot of the Word Cloud population-level list of UMLS concepts associated with orthostatic hypotension. Concepts are listed in descending order by the strength-of-association ratio, that is, the ratio of actual to expected number of cases in the VUMC EHR with the pairwise association of UMLS concepts. The population frequency of each term is also displayed. The ratio is used to rank the condition flowchart. CUI: concept unique identifier; EHR: electronic health record; Misc.: miscellaneous; MsgBskts: message baskets; NewRes: new results; Pop. Freq.: population frequency; Pt.Chart: patient chart; Pt.Lists: patient lists; StNotes: Star Notes; UMLS: Unified Medical Language System; VUMC: Vanderbilt University Medical Center; WhBoards: white boards.

Pt.Chart | StNotes | Forms | Panels | Pt.Lists | TaskList | MsgBskts | WhBoards | NewRes | SignDrafts | Misc.

= | 1 | 2 | 3

Enter one or more UMLS concepts to see population-level associations for all patients with those concepts documented. Ctrl-click o

Term: Show Use a panel

orthostatic hypotension ×

C0020651 71996 orthostatic hypotension

Total: 71996 patients

Filter by type: - All - system: - All -

CUI	Name	Expect	Actual	Ratio	Pop. Freq.
C4076686	supine hypertension	49.565	2721	54.898	2916
C0393911	autonomic failure	50.006	2698	53.953	2942
C1868528	autonomic hypotension	2.363	125	52.907	139
C2315246	pandysautonomia	0.714	36	50.428	42
C0877382	delayed renal graft function	0.544	26	47.801	32
C3662086	multisystem degeneration of autonomic nervous system	0.629	30	47.702	37
C0037019	shy-drager syndrome	8.686	410	47.204	511
C1959798	baroreflex failure syndrome	5.337	237	44.405	314
C0393571	multisystem atrophy	14.482	631	43.572	852
C0877157	pancreas transplant rejection	1.377	56	40.674	81
C5395071	autoimmune ganglionopathy	0.153	6	39.222	9
C1963757	dopamine dysregulation syndrome	0.561	21	37.439	33
C0920233	pisa syndrome	1.122	40	35.656	66

The Word Cloud also aids in identifying concepts most applicable to guide the ranking by displaying each term's UMLS semantic type. In the UMLS Metathesaurus, each concept term is assigned to 1 or more of 127 types in the vocabulary's hierarchical semantic network [20]. Semantic types most relevant for comparison with the condition flowchart include disease or syndrome, injury or poisoning, mental or behavioral dysfunction, sign or symptom, finding, and congenital abnormality. The Word Cloud provides a filter to exclude concepts with semantic types nonrelevant to this task (eg, procedures).

The Word Cloud often lists multiple UMLS concepts that can be grouped to correspond with a single term on the condition flowchart. For example, the Word Cloud concepts associated with orthostatic hypotension include Shy-Drager syndrome, multisystem degeneration of autonomic nervous system, and multisystem atrophy (Figure 1). In 1998, Shy-Drager syndrome was newly categorized as a multisystem atrophy and is no longer the preferred term [21]; the UMLS also lists it as a narrower concept of the term "multiple system atrophy" [8]. In the UMLS, the relationship between "multiple system atrophy" and "multisystem degeneration of autonomic nervous system" is vaguely and imprecisely defined as an "RO" relationship type. RO relationships are described as "other than synonymous, narrower, or broader," however, in this case, the RO determination in the UMLS lacks the relationship attribute that is normally included [8]. The phecodeX map, the term mapping table used for the CKM collaborator's PheWAS, matches "multisystem atrophy" to the phecode "multi-system degeneration of the autonomic nervous system" [22]. Given the evolution of the Shy-Drager syndrome terminology, the UMLS, and the phecodeX mapping, we subsequently considered all 3 of the Word Cloud concepts as a group of related terms; the highest ratio within the group was then used to rank order the condition flowchart.

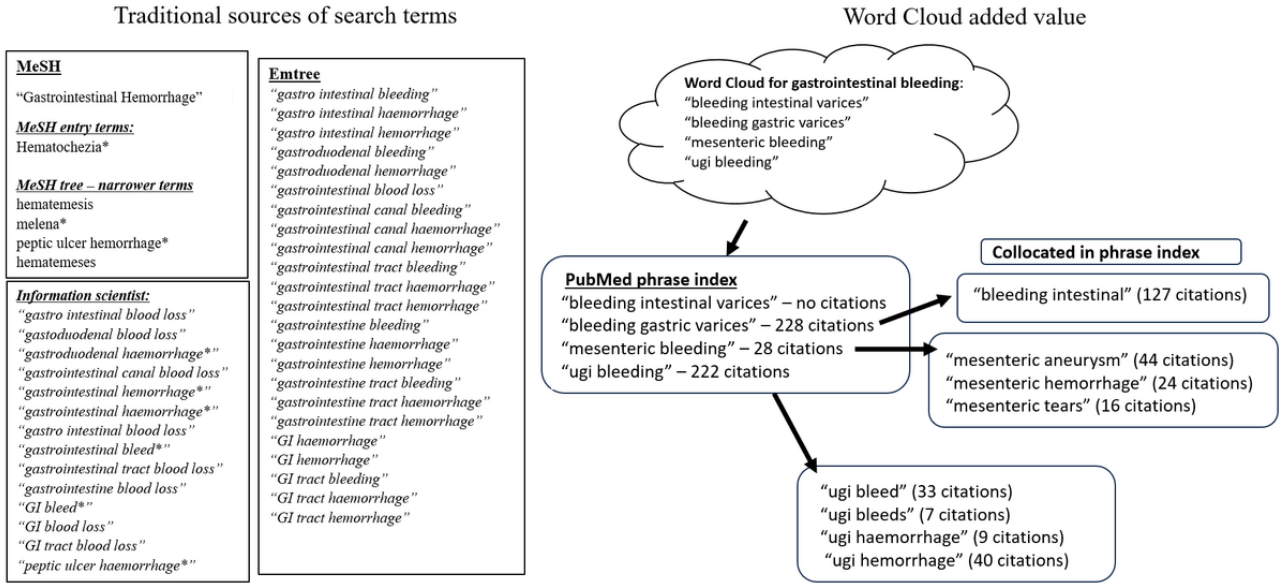
Through the combined processes of documenting actual-to-expected case ratios of the Word Cloud's relevant UMLS concepts, excluding nonrelevant semantic types, and grouping related concepts, our team creates rank-ordered lists of disease causes, risk factors, and complications reflecting our medical center's real-world clinical data.

Substantiating Disease-Condition Associations With Evidence

Providing primary literature to substantiate the associations on the condition flowchart is a key component of our research collaboration. CKM information scientists derive synonyms from the Word Cloud to strengthen the search strategy. When conducting a search, they first compile controlled vocabulary and synonyms from Medical Subject Headings and Emtree [23,24]. Next, they brainstorm additional permutations and extract terms from a scan of the literature; these keywords are subsequently checked for inclusion in the PubMed phrase index.

Figure 2 shows an example of this process for the UMLS concept "gastrointestinal bleeding." Consulting the Word Cloud identified 4 phrases that were not in the initial list of search strategy terms; 3 were found in the PubMed phrase index. Additional terminology was found by scanning collocated terms in the phrase index.

Figure 2. Word Cloud concepts leading to supplemental terminology for a search strategy on gastrointestinal hemorrhage. An asterisk denotes the truncation of a term or phrase to capture permutations. GI: gastrointestinal; MeSH: Medical Subject Headings; UGI: upper gastrointestinal.



In addition to aiding with identifying terminology and concepts to build upon our search strategies, we increasingly realize the importance of the Word Cloud's actual-to-expected patient case ratio for locating appropriate evidence. When creating the condition flowchart, our team may encounter associations for which it is difficult to locate substantiating evidence. In these cases, a Word Cloud ratio that is nonexistent, or lower than 15, can aid in validating literature scarcity. For example, snake bites can lead to nonseptic distributive shock [25]. In the Word Cloud, the association between snake bite and distributive shock has a low ratio of 5.38. Substantiating evidence for the association was found only in case reports and case series (ie, studies with few patients). Similarly, searching for literature to support hypertrophic cardiomyopathy as a cause of obstructive shock yielded only case reports as the best available evidence. A review of the Word Cloud UMLS concepts revealed a ratio of 8.7. In these instances, the evidence may still be used, but the low ranking, due to the low ratio, aids in understanding the strength of association when compared with other causes, risk factors, and complications listed on the condition flowchart.

Conclusions

This viewpoint paper describes a novel use of an institution's AI-driven, large-scale aggregation of condition-specific patient data extracted from free-text clinical documents. The Word Cloud NLP system can inform and guide knowledge generation processes by enhancing our ability to represent, substantiate, and prioritize condition associations for use in PheWAS interpretation.

The VUMC Word Cloud NLP is a valuable resource that provides real-time concept extraction from all clinical documentation and makes the resulting data viewable interactively, available for real-time decision support and alerting, and available as a rich source of coded data for research. An important limitation, however, is that this type of resource would be expensive and difficult to port directly to other institutions, thus limiting its generalizability. The emergence of generative AI, and in particular large language models, makes it conceivable that some of these limitations might be reduced in the near future; for example, large language models might be used to perform a significant portion of the concept extraction task, turning clinical free text into sets of terms which might then be mapped to coded terminologies (such as the UMLS). This possibility is still largely hypothetical and will need to be investigated to evaluate whether it is feasible, performant, and economically viable.

It is also worth noting that in addition to the population-level analysis features offered by the Word Cloud as described in our research collaboration for PheWAS analysis, the CKM also uses its capability of providing summary views of individual patient charts for other projects, such as our synthesized evidence provision services [14,26]. In response to providers' complex clinical questions, information scientists consult the visual display of the Word Cloud to gain a holistic understanding of each patient's comorbidities, medications, and other prominent clinical history. This greatly facilitates our ability to generate tailored syntheses of the published evidence that are personalized to each specific patient case [26]. Additional applications of the Word Cloud and other AI tools are also under exploration at our center, including the use of AI for scaling the maintenance of evidence syntheses over time [27-29]. Through both of these approaches—leveraging the Word Cloud NLP for population-level concept analysis and individual patient-level assessment—the CKM achieves the rapid knowledge acquisition strategy critical for informing clinical health care and research at our institution.

Acknowledgments

This project did not receive any specific external funding.

Authors' Contributions

DAG and NBG wholly developed the work's concept and design. TYK, DAG, AMW, and PAK contributed to the conduct of the case study methods. TYK, DAG, AMW, MNB, PAK, JS, and NBG participated in the writing, editing, and critical review of this paper. AMW and MNB helped visualize the case study details. NBG provided oversight of case study activities.

Conflicts of Interest

None declared.

References

1. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med 2023;155:106649. [doi: 10.1016/j.compbiomed.2023.106649] [Medline: 36805219]
2. Robinson PN, Haendel MA. Ontologies, knowledge representation, and machine learning for translational research: recent contributions. Yearb Med Inform 2020;29(1):159-162 [FREE Full text] [doi: 10.1055/s-0040-1701991] [Medline: 32823310]
3. Giuse DA. Supporting communication in an integrated patient record system. AMIA Annu Symp Proc 2003;2003:1065 [FREE Full text] [Medline: 14728568]
4. Johnson KB, Ehrenfeld JM. An EPIC switch: preparing for an electronic health record transition at Vanderbilt University Medical Center. J Med Syst 2017;42(1):6 [FREE Full text] [doi: 10.1007/s10916-017-0865-6] [Medline: 29164347]
5. Madani S, Giuse D, McLemore M, Weitkamp A. Augmenting NLP results by leveraging SNOMED CT relationships for identification of implantable cardiac devices from patient notes. : SNOMED International; 2019 Presented at: SNOMED CT Expo; Oct 31-Nov 1, 2019; Kuala Lumpur, Malaysia URL: http://tinyurl.com/5awcneyr
6. Tan H, Giuse D, Kumah-Crystal Y. Usability of a word cloud visualization of the problem list. Washington, DC: American Medical Informatics Association; 2020 Presented at: AMIA Clinical Informatics Conference; May 19-21, 2020; Virtual URL: https://brand.amia.org/m/1b1f63ea67b0c099/original/CIC2020-Visual-Abstract-Collection-FINAL.pdf
7. Krause KJ, Shelley J, Becker A, Walsh C. Exploring risk factors in suicidal ideation and attempt concept cooccurrence networks. AMIA Annu Symp Proc 2023;2022:644-652 [FREE Full text] [Medline: 37128429]
8. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
9. Apache cTAKES. Apache Software Foundation. URL: https://ctakes.apache.org/ [accessed 2023-12-13]
10. MetaMap speed. National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/Docs/Speed.pdf [accessed 2023-12-13]
11. Albert D, Weitkamp AO, Giuse D, Wright A. Building a pipeline for clinical alerts generated via natural language processing. 2022 Presented at: American Medical Informatics Association Annual Symposium; 2022 Nov; Washington, DC URL: https://knowledge.amia.org/event-data/esearch?q=Building%20a%20pipeline%20for%20clinical%20alerts%20generated%20via%20natural%20language%20processing&size=n_20_n
12. Giuse NB, Kusnoor SV, Koonce TY, Ryland CR, Walden RR, Naylor HM, et al. Strategically aligning a mandala of competencies to advance a transformative vision. J Med Libr Assoc 2013;101(4):261-267 [FREE Full text] [doi: 10.3163/1536-5050.101.4.007] [Medline: 24163597]
13. Giuse NB, Koonce TY, Jerome RN, Cahall M, Sathe NA, Williams A. Evolution of a mature clinical informationist model. J Am Med Inform Assoc 2005;12(3):249-255 [FREE Full text] [doi: 10.1197/jamia.M1726] [Medline: 15684125]
14. Blasingame MN, Williams AM, Su J, Naylor HM, Koonce TY, Epelbaum MI, et al. Bench to bedside: detailing the catalytic roles of fully integrated information scientists. Mount Laurel, NJ: Special Libraries Association; 2019 Presented at: Special Libraries Association Annual Meeting; June 18, 2019; Cleveland, OH URL: https://www.sla.org/learn-2/research/sla-contributed-papers/2019-contributed-papers/
15. Giuse NB, Williams AM, Giuse DA. Integrating best evidence into patient care: a process facilitated by a seamless integration with informatics tools. J Med Libr Assoc 2010;98(3):220-222 [FREE Full text] [doi: 10.3163/1536-5050.98.3.009] [Medline: 20648255]
16. Stead WW, Lewis A, Giuse NB, Koonce TY, Bastarache L. Knowledgebase strategies to aid interpretation of clinical correlation research. J Am Med Inform Assoc 2023;30(7):1257-1265. [doi: 10.1093/jamia/ocad078] [Medline: 37164621]
17. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;26(9):1205-1210 [FREE Full text] [doi: 10.1093/bioinformatics/btq126] [Medline: 20335276]

18. Bastarache L, Denny JC, Roden DM. Phenome-wide association studies. JAMA 2022;327(1):75-76 [FREE Full text] [doi: 10.1001/jama.2021.20356] [Medline: 34982132]
19. Bland M. An Introduction to Medical Statistics, 4th Edition. Oxford, UK: Oxford University Press; 2015.
20. Semantic network. UMLS® Reference Manual. Bethesda, MD: National Library of Medicine; 2009. URL: https://www.ncbi.nlm.nih.gov/books/NBK9679/ [accessed 2023-12-13]
21. Fecek C, Nagalli S. Shy-Drager Syndrome. Treasure Island, FL: StatPearls Publishing; 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK560502/ [accessed 2023-12-13]
22. PhecodeX (Extended). PheWAS Resources. URL: https://phewascatalog.org/phecode_x [accessed 2023-12-13]
23. Medical Subject Headings (MeSH). National Library of Medicine. 2023. URL: https://www.nlm.nih.gov/mesh/meshhome.html [accessed 2023-12-13]
24. Emtree. Elsevier. URL: https://www.elsevier.com/products/embase/emtree [accessed 2024-01-10]
25. Gaieski DF, Mikkelsen ME. Definition, classification, etiology, and pathophysiology of shock in adults. UpToDate. 2023. URL: https://www.uptodate.com/contents/definition-classification-etiology-and-pathophysiology-of-shock-in-adults [accessed 2023-12-13]
26. Koonce TY, Giuse DA, Blasingame MN, Su J, Williams AM, Biggerstaff PL, et al. The personalization of evidence: using intelligent datasets to inform the process. Washington, DC: American Medical Informatics Association; 2020 Presented at: AMIA Fall Symposium; November 2020; Virtual.
27. Koonce TY, Blasingame MN, Williams AW, Clark JD, DesAutels SJ, Giuse DA, et al. Building a scalable knowledge management approach to support evidence provision for precision medicine. Washington, DC: American Medical Informatics Association; 2022 Presented at: AMIA Informatics Summit; March 2022; Chicago, IL.
28. Blasingame MN, Su J, Zhao J, Clark JD, Koonce TY, Giuse NB. Using a semi-automated approach to update clinical genomics evidence summaries. Chicago, IL: Medical Library Association; 2023 Presented at: Medical Library Association and Special Libraries Association Annual Meeting; May 2023; Detroit, MI.
29. Su J, Blasingame MN, Zhao J, Clark JD, Koonce TY, Giuse NB. Using a performance comparison to evaluate four distinct AI-assisted citation screening tools. Chicago, IL: Medical Library Association; 2024 Presented at: Medical Library Association Annual Meeting; May 2024; Portland, OR.

Abbreviations

AI: artificial intelligence
CKM: Center for Knowledge Management
EHR: electronic health record
NLP: natural language processing
PheWAS: phenome-wide association study
UMLS: Unified Medical Language System
VUMC: Vanderbilt University Medical Center

Edited by G Eysenbach, C Lovis; submitted 10.10.23; peer-reviewed by P Han, D Chrimes; comments to author 08.12.23; revised version received 15.12.23; accepted 04.01.24; published 30.01.24.

Please cite as:
Koonce TY, Giuse DA, Williams AM, Blasingame MN, Krump PA, Su J, Giuse NB
Using a Natural Language Processing Approach to Support Rapid Knowledge Acquisition
JMIR Med Inform 2024;12:e53516
URL: https://medinform.jmir.org/2024/1/e53516
doi:10.2196/53516
PMID:38289670

Viewpoint

The Current Status and Promotional Strategies for Cloud Migration of Hospital Information Systems in China: Strengths, Weaknesses, Opportunities, and Threats Analysis

Jian Xu[1], MSc, MPH

Department of Health Policy, Beijing Municipal Health Big Data and Policy Research Center, Beijing, China

Corresponding Author:
Jian Xu, MSc, MPH
Department of Health Policy
Beijing Municipal Health Big Data and Policy Research Center
Building 1, Number 6 Daji Street
Tongzhou District
Beijing, 101160
China
Phone: 86 01055532146
Email: xujian@163.com

Abstract

Background: In the 21st century, Chinese hospitals have witnessed innovative medical business models, such as online diagnosis and treatment, cross-regional multidepartment consultation, and real-time sharing of medical test results, that surpass traditional hospital information systems (HISs). The introduction of cloud computing provides an excellent opportunity for hospitals to address these challenges. However, there is currently no comprehensive research assessing the cloud migration of HISs in China. This lack may hinder the widespread adoption and secure implementation of cloud computing in hospitals.

Objective: The objective of this study is to comprehensively assess external and internal factors influencing the cloud migration of HISs in China and propose promotional strategies.

Methods: Academic articles from January 1, 2007, to February 21, 2023, on the topic were searched in PubMed and HuiyiMd databases, and relevant documents such as national policy documents, white papers, and survey reports were collected from authoritative sources for analysis. A systematic assessment of factors influencing cloud migration of HISs in China was conducted by combining a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis and literature review methods. Then, various promotional strategies based on different combinations of external and internal factors were proposed.

Results: After conducting a thorough search and review, this study included 94 academic articles and 37 relevant documents. The analysis of these documents reveals the increasing application of and research on cloud computing in Chinese hospitals, and that it has expanded to 22 disciplinary domains. However, more than half (n=49, 52%) of the documents primarily focused on task-specific cloud-based systems in hospitals, while only 22% (n=21 articles) discussed integrated cloud platforms shared across the entire hospital, medical alliance, or region. The SWOT analysis showed that cloud computing adoption in Chinese hospitals benefits from policy support, capital investment, and social demand for new technology. However, it also faces threats like loss of digital sovereignty, supplier competition, cyber risks, and insufficient supervision. Factors driving cloud migration for HISs include medical big data analytics and use, interdisciplinary collaboration, health-centered medical service provision, and successful cases. Barriers include system complexity, security threats, lack of strategic planning and resource allocation, relevant personnel shortages, and inadequate investment. This study proposes 4 promotional strategies: encouraging more hospitals to migrate, enhancing hospitals' capabilities for migration, establishing a provincial-level unified medical hybrid multi-cloud platform, strengthening legal frameworks, and providing robust technical support.

Conclusions: Cloud computing is an innovative technology that has gained significant attention from both the Chinese government and the global community. In order to effectively support the rapid growth of a novel, health-centered medical industry, it is imperative for Chinese health authorities and hospitals to seize this opportunity by implementing comprehensive strategies aimed at encouraging hospitals to migrate their HISs to the cloud.

(JMIR Med Inform 2024;12:e52080) doi:10.2196/52080

KEYWORDS

hospital information system; HIS; cloud computing; cloud migration; Strengths, Weaknesses, Opportunities, and Threats analysis

Introduction

In the 21st century, innovative business models have emerged in Chinese hospitals, such as online diagnosis and treatment, cross-regional multidepartment consultation, real-time sharing of medical test results, and continuous public health surveillance. However, most hospitals still rely on traditional hospital information systems (HISs) designed for in-hospital management that are inadequate to support the development of these new business models [1]. Cloud computing has emerged as a promising global information technology recognized as a new infrastructure for future economic growth [2,3]. Since 2010, it has also been prioritized by the Chinese government as a "national strategic emerging industry" [4]. The adoption of cloud computing technologies can significantly reduce hospitals' costs associated with system construction and maintenance [5], expand medical services to partner institutions or patients outside the hospital [6], provide more secure network protection than self-built data centers [7], and facilitate large-scale collection and analysis of clinical data essential for scientific clinical decision-making [8]. Based on these advantages, there has been a surge in China's medical cloud service market and application research in recent years [9].

However, despite the increased attention given to cloud computing in various disciplinary domains such as disease monitoring, health surveillance, and clinical diagnosis, there is a lack of research on the cloud migration of HISs. A comprehensive review of the PubMed and HuiyiMd databases only yielded 3 relevant studies: an Iranian study that identified key driving factors for hospitals adopting cloud computing [10], a Greek study that proposed a method for migrating clinical and laboratory data based on local hospital conditions [11], and an American study that focused on essential considerations for chief financial officers before venturing into the cloud [12]. However, none of these studies have adequately addressed the aforementioned issue. Without conducting prior assessments, hospitals may struggle to fully comprehend the external environment, internal conditions, and potential opportunities and risks, thus failing to ensure prudent decision-making. Blindly following trends could pose significant threats to the security, operational efficiency, and maintenance costs of already deployed cloud-based information systems and existing hospital networks [13]. Therefore, this study aims to systematically assess factors influencing the cloud migration of HISs in China, identify associated challenges, and propose corresponding strategies for advancement. It can assist hospitals in gaining a comprehensive understanding of this work while safely implementing their cloud-based medical services. Additionally, it serves as a foundation for formulating policies aligned with Chinese hospital informatization development in the new era by health authorities while being referenced by other countries or regions facing similar challenges.

Methods

Information Sources

The primary data source for this study was obtained from literature databases to understand the practical applications of cloud technology in Chinese hospitals. The articles published between January 1, 2007, and February 21, 2023, were selected from MEDLINE (accessible through PubMed) and HuiyiMd (accessible through the Huiyi Medical Literature Express Service System).

In order to overcome the inherent limitations of academic articles, this study augmented a wealth of pertinent internal and external environmental information by extensively consulting authoritative sources such as government agencies, industry organizations, academic institutions, and market research companies. These sources of information included national policies, action plans, white papers, implementation guidelines, survey reports, and statistical data from the past 10 years.

Search Strategies

The search strategy for PubMed: (((cloud [Title/Abstract]) OR (cloud-based [Title/Abstract])) AND (hospital [Title/Abstract]) AND ("2007/01/01" [Date-Publication]: "2023/02/21" [Date-Publication])). The search strategy for HuiyiMd: ([TI] (cloud AND hospital) OR [AB] (cloud AND hospital) OR [MH] (cloud AND hospital)) AND ([PY]>=2007). The search strategy for authoritative sources involves entering the keywords "hospital" AND "cloud" in the search box on websites.

Inclusion and Exclusion Criteria

Based on specified inclusion and exclusion criteria (Textbox 1), irrelevant articles or those covering the same topic from the same institution were excluded. Subsequently, an Excel (Microsoft) spreadsheet (Multimedia Appendix 1) and a reference list (Multimedia Appendix 2 [1,2,6,8,14-26]) were generated for literature review and Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis.

Textbox 1. Inclusion and exclusion criteria for literature review.

Inclusion criteria

- Article type: fully retrievable
- Language: English, Chinese
- Nationality of the first author: Chinese (including Hong Kong and Taiwan)
- Article topic: the research, development, and application of cloud technology in Chinese hospitals
- Publication date: from January 1, 2007, to February 21, 2023

Exclusion criteria

- Article type: nonretrievable
- Language: other languages
- Nationality of the first author: other countries
- Article topic: other topics
- Publication date: before January 1, 2007; after February 21, 2023

Information Extraction

The accessible articles were assessed based on the following criteria: title, authors, first author, first author affiliation, publication year, journal name, digital object identifier (DOI), PubMed unique identifier (PMID), first author nationality, abstract, and conclusion. Furthermore, the positive and negative impacts, research methods, disciplinary domains, cloud service models, and institutional affiliations were taken into account for further in-depth analysis purposes. The findings were documented and statistically analyzed in Excel.

Analysis Methods

The SWOT analysis is a systematic assessment of strengths (S), weaknesses (W), opportunities (O), threats (T), and other factors that influence a specific topic, objectively describing the current situation of an organization or enterprise and formulating corresponding strategies [14]. It is widely used in strategic decision-making and competitor analysis within organizations or businesses due to its ability to simplify complex problems into essential issues, enabling more focused problem-solving. This study uses the SWOT method to assess the factors impacting China's cloud migration of HISs and proposes promotional strategies.

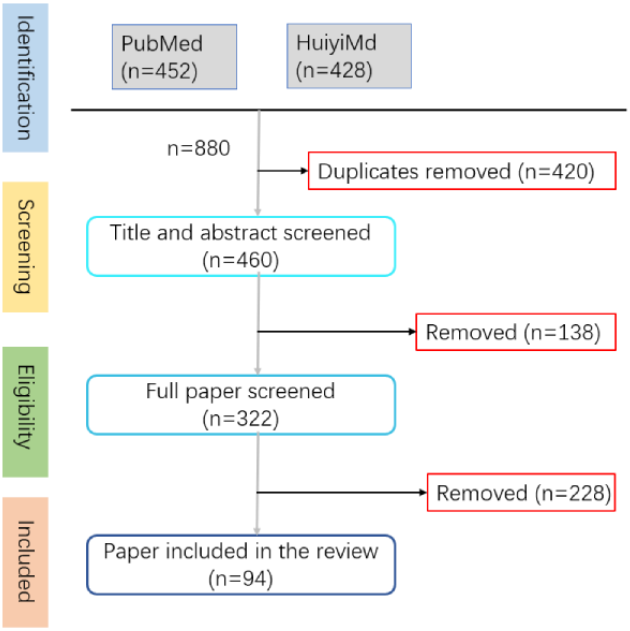
Results

Literature Review

Identification Process

The identification process in this study consists of four steps (Figure 1): (1) A total of 880 articles were retrieved from PubMed and HuiyiMd databases. (2) The search results were amalgamated, resulting in 460 deduplicated articles. (3) Screening the titles and abstracts eliminated 138 irrelevant articles based on the exclusion criteria. (4) The full text of the remaining articles was meticulously examined against predefined inclusion and exclusion criteria, resulting in a final selection of 94 relevant articles.

Figure 1. Flow diagram for the identification process.

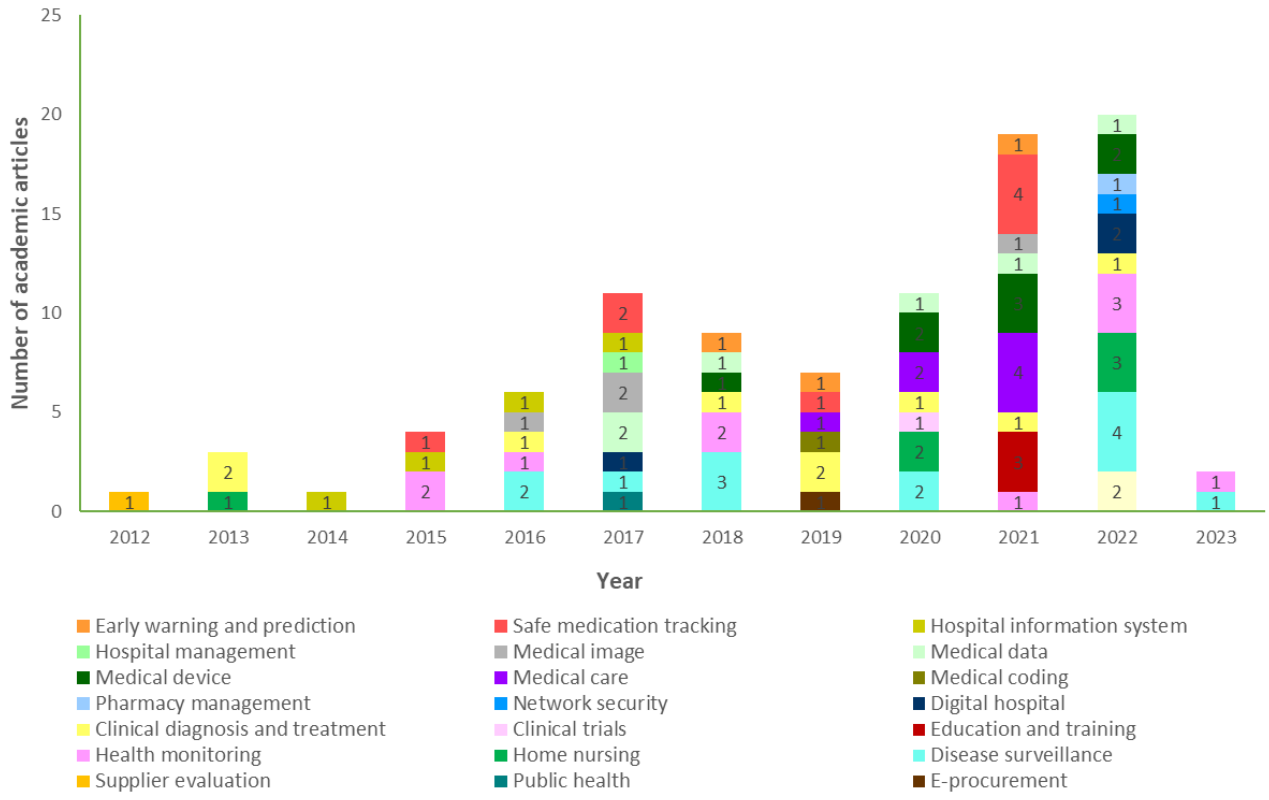


Comprehensive Description of the Literature

Research and Application of Cloud Technology in Hospitals Has Grown Rapidly and Continuously Expanded in Disciplinary Domains

In terms of time line, there was a gradual step-like increase in the number of articles starting in 2012 and reaching its peak at 20 articles in 2022. The compound annual growth rate (CAGR) was approximately 35%, highlighting the escalating quantity of research into and application of cloud technology in hospitals, as shown in Figure 2. The number of disciplines involved has increased from 1 in 2012 to 10 in 2022, encompassing a total of 22 domains. Specifically, research and application are predominantly observed in the domains of disease monitoring, health surveillance, clinical diagnosis and treatment, safe medication tracking, and medical devices, constituting 51% (48/94) of the overall distribution.

Figure 2. Time distribution of disciplinary domains involved in academic articles.



Implementation of Cloud Technology Can Yield Favorable Outcomes for Hospitals to a Certain Extent

The analysis of 94 articles identified 3 categories and 9 research methods (Table 1). The "technology" category was the most prevalent, with 47 (50%) articles focusing on information systems, cloud platforms, and associated technologies. The "experience" category followed closely, with 40 (43%) articles, primarily validating the performance of or applying cloud-based information systems through empirical research, case-control studies, experience sharing, and cohort studies. Finally, the "literature" category consisted of only 7 literature reviews on this subject matter. The consistent findings of these studies demonstrate the implementation of cloud technology in hospitals can yield positive impact to some extent, such as enhancing precision in management practices, expanding disease monitoring capabilities, reducing workload for medical personnel, and providing convenient and cost-effective health care services for patients. However, 5% (n=5) of the articles also acknowledged certain negative impacts, such as underdevelopment of digital methods in hospitals, cybersecurity risks, and low satisfaction rates among physicians and community pharmacists.

Table 1. The correlation between research methods used in academic articles and the institutional affiliations of their first authors.

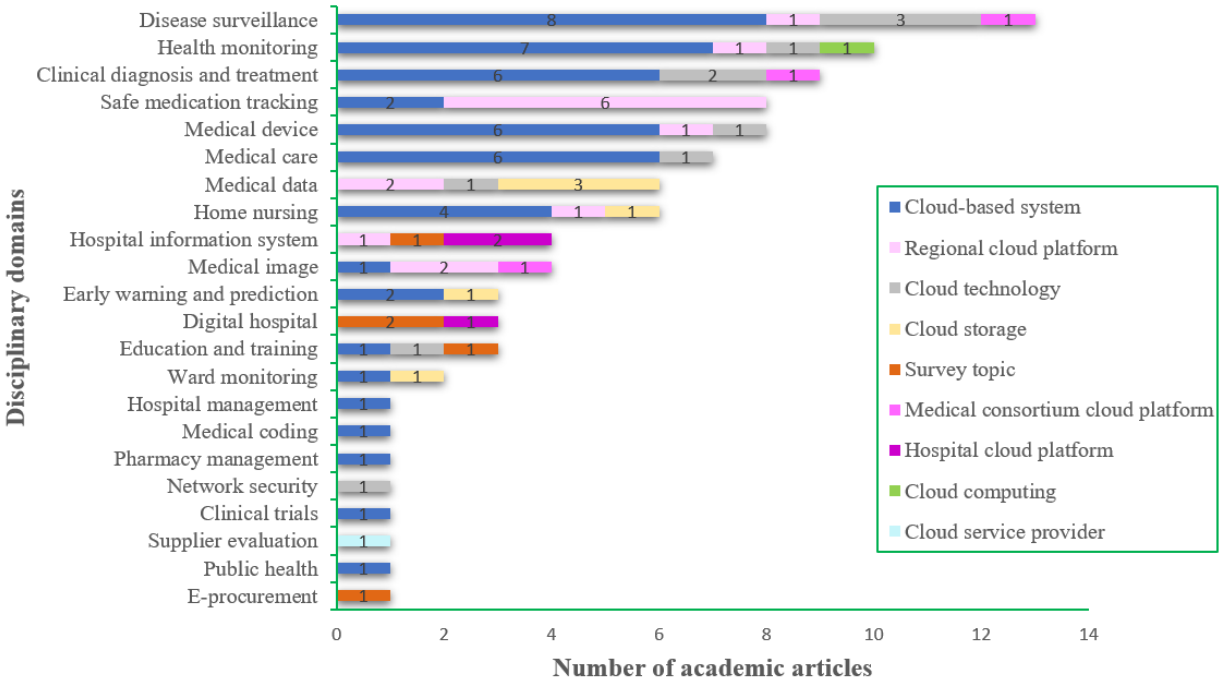
Research methods	Hospitals, n (%)	Universities or colleges, n (%)	Associations or companies, n (%)
Technology			
System research and development	13 (14)	14 (15)	N/A[a]
Cloud platform construction	5 (5)	5 (5)	N/A
Technical research	2 (2)	8 (9)	N/A
Experience			
Empirical research	14 (15)	7 (7)	1 (1)
Case control study	11 (12)	2 (2)	N/A
Summary of experience	3 (3)	N/A	N/A
Cohort study	2 (2)	N/A	N/A
Literature			
Retrospective study	4 (4)	2 (2)	N/A
Standard study	N/A	1 (1)	N/A

[a]N/A: not applicable.

More Than Half of the Studies Focused on Task-Specific Cloud-Based Systems, While Only 1 in 5 Addressed Integrated Cloud Platforms

Out of the 94 articles analyzed, the majority (n=49, 52%) focused on task-specific cloud-based systems in hospitals. In contrast, only 21 (22%) articles discussed or developed integrated cloud platforms for sharing within a region, medical alliance, or hospital. Furthermore, as shown in Figure 3, 67% (n=33) of task-specific cloud-based systems were used in patient-related domains such as disease monitoring, health surveillance, clinical diagnosis and treatment, medical care, and medical devices. A total of 15 regional cloud platforms (16%) were commonly used for safe medication tracking, data storage, and medical imaging. A total of 3 medical alliance cloud platforms (3%) found use in disease treatment-related domains such as disease monitoring, clinical diagnosis, and treatment along with medical imaging. A total of 3 hospital cloud platforms (3%) primarily originated from digital hospitals or HIS upgrades.

Figure 3. The cross-distribution of cloud service models and disciplinary domains covered in academic articles.



More Supplementary Materials Were Collected to Support the SWOT Analysis for This Study

Because the literature review provided insufficient information to support the analysis of internal and external factors for this study, supplementary materials were collected, including national policies, action plans, white papers, implementation guidelines, research reports, and statistical reports from authoritative websites, such as government agencies, industry organizations, academic institutions, and market research companies. In total, 37 supplementary documents were included in the SWOT analysis: 4 policy papers, 11 industry reports, 15 academic articles, 2 dissertations, 2 official bulletins, and 3 news articles. All of them were recorded in an Excel file (Multimedia Appendix 2).

SWOT Analysis

External Opportunities (O)

Politics: Governments Worldwide Prioritize Cloud Technology and Have Implemented Supportive Policies

The United States introduced the Cloud First policy [2] and the CLOUD Act [15]. Similarly, the European Union aims for digital sovereignty through initiatives like the Gaia-X Association and the EU Cloud Computing Strategy [27]. China also prioritized cloud computing as a "national strategic emerging industry" in 2010 and implemented policies to promote its adoption by the government and businesses [28]. Consequently, China has rapidly developed in this field and ranks among global leaders [29].

Economy: Both Nations and Enterprises Have Made Substantial Capital Investments to Foster Its Development

According to Gartner's data from 2011 to 2022, the market scale increased from US $95.24 billion to US $491 billion with a CAGR exceeding 16% [30]. The global medical cloud computing market reached US $39.4 billion, with Asia-Pacific exhibiting the fastest growth rate at 22% per year; China and India are significant contributors to this expansion [9]. In China, the market size has surged from less than US $270 million in 2011 to US $66.91 billion in 2022, with a consistent CAGR surpassing 40%, and will exceed US $150 billion by 2025 [30].

Social: Online Health Care Has Become a Norm in Modern Life

According to the National Telemedicine and Connected Healthcare Center of China, as of June 2021, there were 239 million users accessing health care services online and more than 1600 internet-based hospitals in China [31]. Another survey revealed that approximately 63% (n=465) of the surveyed hospitals (738 hospitals across 30 provinces) used cloud services to some extent in 2022 [32]. Moreover, the COVID-19 pandemic has further fueled the demand for online health care services [6]. Consequently, these services have become as commonplace in our lives as online shopping.

Older Adult Care: The Older Adult Care Industry Urgently Needs Advanced Technological Support

By the end of November 2020, China's population of people aged 60 years and older was 264,018,766, accounting for 19% of the total population, making it the country with both the largest older population and the fastest-aging society worldwide [33]. Consequently, relying solely on their children, nursing homes, or communities to provide older adult care services has become increasingly impractical. Recognizing this challenge, the National Health Commission of China issued a document in October 2019 emphasizing the need to fully harness modern technologies such as cloud computing, artificial intelligence (AI), and the Internet of Things to develop an intelligent service model known as "Internet plus Healthy Aging" [34].

External Threats (T)

Market: Technological Monopolies Pose a Significant Threat to the Digital Autonomy of Nations

A total of 81 (81%) of the Forbes Top 100 cloud computing companies are American [35], and they possess significant technological and capital advantages. They continuously expand their global market share; they captured 60%-70% (JPY ¥1,725-¥2,012 billion [US $11.6-$13.6 million]) of the Japanese cloud market in 2020 [36] and 69% of the European cloud market in 2021 [37]. The passage of the Clarifying Lawful Use of Data Abroad Act (CLOUD Act) by the US Congress in March 2018 caused European countries to feel threatened due to the potential loss of digital sovereignty [15], which prompted them to initiate their "European cloud" project in 2020 [27]. Other nations may face similar challenges.

Competition: Fierce Competition May Lead to Uncertain Levels of Service Quality

In November 2021, 5 bidders for a public cloud service project in Shijiazhuang City, China, submitted bids of CNY ¥0, sparking concerns among stakeholders [38]. The intense competition may lead to unpredictable service performance issues for users, such as limiting hospitals' access to better pricing and a wider range of choices by restricting the interoperability and portability of HISs or causing sudden disruptions in cloud-based medical services after winning the bid, which poses significant risks to patient safety [16].

Security: Hospitals Express Apprehensions Regarding Diverse Cyber-Attacks Targeting Cloud Infrastructure

Security is the primary challenge for cloud-based systems due to various cyberattacks faced by current cloud environments, especially in the health care sector where sensitive data such as personal privacy, health records, diagnostics, and treatments are stored [13]. Even prominent cloud providers like Azure (Microsoft), Docker Hub (Docker), and Everis (NTT DATA) have experienced malicious intrusions [30], while both the United Kingdom's National Health Service (NHS) in 2017 and Ireland's Department of Health information system in 2021 were both targeted and resulted in a complete paralysis [39,40].

Legislation: The Lack of Precise Legislation Hinders the Efficient Implementation and Enforcement of Regulatory Measures

To support the implementation of the "Cloud Normal" and "Internet Plus" strategies, the Chinese government has enacted laws, regulations, and management measures. However, there are limited directly applicable legal provisions for cloud migration of HISs. Imperfect laws and regulations, insufficient safety standards, unclear legal liabilities, and the absence of a damage assessment mechanism hinder the proper development of cloud services. As a result, doctors and patients may encounter challenges in protecting their rights during disputes [17].

Internal Strengths (S)

Data: Hospitals Generate Substantial Volumes of Medical Data on a Daily Basis

Hospitals are natural suppliers of big data. For instance, the Chinese National Cloud-Based Telepathology System (CNCTPS) has collected 23,167 cases and served 9240 users in 4 years (2016-2019), providing comprehensive details from whole-slide images to diagnostic reports [5]. Additionally, medical big data can provide substantial value to both hospitals and patients. For example, the aforementioned CNCTPS application can save patients around US $300,000 per year [5]. Abundant electronic health and care records in the United Kingdom's NHS can reduce hospital operational costs by approximately £5 billion (US $ 3.9 billion) annually and save patient welfare expenses by roughly £4.6 billion (US $3.6 billion) [41]. Furthermore, traditional computing tools are unable to handle the processing and analysis of such massive amounts of data—this is exactly where cloud computing technology excels [18].

Business: The Provision of Comprehensive Medical Services Necessitates Extensive Interdisciplinary Collaboration

Medical services are complex and innovative, requiring synchronization of knowledge, technology, experience, and resources from diverse disciplines. Cloud computing provides extensive connectivity, offering robust support for these tasks, including interdisciplinary expert consultations, collaborative surgeries, and integrating medicine and care [19]. The Huashan Hospital, affiliated with Fudan University, uses a medical consortium cloud platform where experts from higher-level hospitals offer diagnostic advice to lower-level hospitals for subsequent care and daily treatment, ensuring positive outcomes for patients with epilepsy [42].

Application: Multiple Cloud Technology Applications Have Been Effectively Implemented Across Various Medical Domains

As shown in Figure 2, cloud technology is receiving increasingly extensive research and application in the medical field, and even some regional or medical alliances have constructed their own medical cloud platforms to store health data, share medical images, and facilitate collaboration. Furthermore, a national survey conducted in 2022 also confirmed these findings by revealing that out of the 738 surveyed hospitals, 63% (n=465) partially used cloud services across nearly 20 different medical business scenarios [32]. These effective practices can serve as valuable references and support for other hospitals yet to implement such initiatives.

Demand: The Provision of Health-Centered Medical Services Necessitates Advanced Technological Support

The transition from disease-centered to health-centered hospital development in the new era has rendered traditional HISs increasingly inadequate as they were previously designed solely for managing information within hospitals. Cloud computing can significantly expand hospitals' medical services beyond their physical premises, enabling online chronic disease management, individual life-cycle health surveillance, and remote diagnosis for patients in remote areas. This enhancement

empowers hospitals to provide health-centered medical services [20]. The findings of this study also strongly support this notion. As depicted in Figure 3, cloud technology has been extensively used in closely associated domains with patients, encompassing disease monitoring, health surveillance, clinical diagnosis and treatment, and safe medication tracking.

Internal Weaknesses (W)

System: The Complexity of HISs Poses Challenges for Hospitals When Migrating Them to the Cloud

The HISs are the most complex organizational information management systems developed by various contractors in diverse environments, covering a wide range of business functions and user groups [21], as depicted in Figure 3, with only 94 articles included but spanning across 22 distinct disciplinary domains as well. Therefore, the cloud migration of HISs presents significant challenges, particularly for those systems abandoned by development companies due to insolvency or insufficient technical support. Nevertheless, if there existed an all-encompassing and authoritative medical cloud platform enabling hospitals to tailor services based on their specific requirements, it would undoubtedly expedite the overall migration process.

Security: The Security of Existing Hospital Networks Still Faces Numerous Risks

Currently, most HISs still operate in self-constructed networks instead of using cloud-based solutions, which poses information security challenges due to insufficient infrastructure, overreliance on a single protective measure, incomplete regulatory frameworks, and potential vulnerabilities from privilege abuse [22]. For example, the 2019-2020 China Hospital Informatization Survey Report revealed that around 28% (n=282) of surveyed hospitals experienced unplanned core system failures lasting more than 30 minutes [43]. To effectively address these concerns, proficient IT teams like reputable cloud vendors or organizations equipped with advanced technologies such as cloud computing should collaborate rather than solely rely on in-house hospital IT capabilities.

Plan: Strategic Planning and Resource Allocation in Hospitals Exhibit Certain Deficiencies

According to Figure 3, more than half of the research articles focused on hospital-specific systems for various tasks. These systems still adhered to traditional information system designs, had limited scalability and functionality, and operated independently. As a result, there were significant challenges in effectively using cloud computing's computing capabilities, storage capacity, and integrated analysis to generate valuable information supporting government scientific decision-making. The survey results from China's National Health Commission also confirmed this point as many internet hospitals were not fully developed yet and encountered diverse issues [44]. Moreover, only 14% (n=101) of hospitals had migrated their core business operations to the cloud with a mere 13% (n=100) planning to do so in the next 3-5 years [32]. Therefore, it is crucial for hospitals to reorganize and integrate their operations and resources before incorporating their HISs into the cloud in order to meet the demands of cloud capabilities and new health care models.

Personnel: The Allocation of Information Personnel Is Inadequate and Lacks Specialization Levels

With the increasing integration of cloud computing, AI, and robotics, hospitals urgently require highly skilled IT professionals to effectively implement these new technologies [23]. However, a national survey in 2021 revealed that the average number of information department personnel in 9376 secondary and tertiary hospitals was only 6. Most of these personnel held undergraduate or junior college computer degrees and possessed limited interdisciplinary expertise. This falls significantly below national standards [24], particularly for hospitals below grade 2 or in economically underdeveloped areas [45].

Investment: Primary Hospitals Lack Sufficient Investment in Information Technology and Cloud Services

According to a 2020 survey by the National Health Commission of China, most primary hospitals allocated less than 1% of their budgets to HIS development, facing challenges such as unstable funding and support [25]. A nationwide survey conducted in 2022 revealed that only 53% of the surveyed 738 hospitals had expenses related to public cloud services in the previous 2 years, with 54% spending less than US $14,000. Particularly for primary hospitals, establishing a cloud service system is even more financially challenging [32]. Clearly, these primary hospitals require more reliable financial guarantees for the smooth operation of their HISs and cloud services.

Discussion

Principal Findings

Extensive literature review and systematic SWOT analysis indicate that cloud computing is increasingly being applied in nearly 22 discipline domains in Chinese hospitals; it plays a crucial role in monitoring patient-related diseases, health surveillance, clinical diagnosis and treatment, and safe medication tracking. However, more than half of the research and applications are limited to cloud-based systems for specific hospital tasks, which fail to fully leverage the robust integrated analytical capabilities of cloud computing due to limited data scale and functionality that could otherwise generate valuable information supporting hospital or government decision-making processes. Additionally, challenges such as market sovereignty disputes, intense industry competition, network attacks, inadequate regulation, and hospitals' internal weaknesses like complexity of HISs, insufficient resource integration, and limited manpower and investment, hinder widespread adoption of cloud technology among most hospitals that exhibit a relatively weak willingness to migrate their core operations to the cloud within the next 3-5 years. Nevertheless, cloud computing is widely recognized as a novel infrastructure driving global economic growth. Integrating cloud technology in hospitals can enhance medical service quality, foster interdisciplinary collaboration and remote consultations, and promote coordinated development within regional health care economies. Consequently, it is imperative for hospitals and health authorities to pay special

attention to this matter and actively implement diverse strategies to facilitate its advancement. Based on the aforementioned research findings, this study proposes a set of promotional strategies for collective deliberation among peers. The overall framework depicting these proposed strategies is illustrated in Figure 4, which will be further elucidated in subsequent sections.

Figure 4. SWOT analysis and response strategies diagram for cloud migration of HISs. HIS: hospital information system; SWOT: Strengths, Weaknesses, Opportunities, and Threats.

Internal factors / External factors	Strengths (S)	Weaknesses (W)
	1. Data: Hospitals generate substantial volumes of medical data on a daily basis. 2. Business: The provision of comprehensive medical services necessitates extensive interdisciplinary collaboration. 3. Application: Multiple cloud technology applications have been effectively implemented across various medical domains. 4. Demand: The provision of health-centered medical services necessitates advanced technological support.	1. System: The complexity of HISs poses challenges for hospitals when migrating them to the cloud. 2. Security: The security of existing hospital networks still faces numerous risks. 3. Plan: The strategic planning and resource allocation in hospitals exhibit certain deficiencies. 4. Personnel: The allocation of information personnel is inadequate and lacks specialization levels. 5. Investment: Primary hospitals lack sufficient investment in information technology and cloud services.
Opportunities (O)	**OS strategy**	**OW strategy**
1. Politics: Governments worldwide prioritize cloud technology and have implemented supportive policies. 2. Economy: Both nations and enterprises have made substantial capital investments to foster its development. 3. Social: Online health care has become a norm in modern life. 4. Older adult care: The older adult care industry urgently needs advanced technological support.	**Implementing multiple initiatives to encourage more hospitals to migrate their HISs to the cloud.** 1. Taking multiple initiatives to encourage more hospitals to migrate their HISs to the cloud. 2. Guiding hospitals in safely and effectively migrating their systems to the cloud.	**Enhancing hospitals' overall capability for cloud migration of HISs.** 1. Improving the medical information literacy and capabilities of all staff in hospitals. 2. Reengineering hospital's business processes and integrating them into one organic HIS. 3. Increasing investment in and application of cloud technology in primary hospitals.
Threats (T)	**TS strategy**	**TW strategy**
1. Market: Technological monopolies pose a significant threat to the digital autonomy of nations. 2. Competition: Fierce competition may lead to uncertain levels of service quality. 3. Security: Hospitals express apprehensions regarding diverse cyber-attacks targeting cloud infrastructure. 4. Legislation: The lack of precise legislation hinders the efficient implementation and enforcement of regulatory measures.	**Establishing a provincial-level unified medical hybrid multicloud platform.** 1. Selecting the most suitable cloud deployment model for hospitals. 2. Establishing a unified medical cloud platform at the provincial or municipal level.	**Enhancing legal framework and technical support for cloud-based HISs.** 1. Establishing a solid and reliable legal foundation. 2. Providing comprehensive and efficient technical support.

Implementing Multiple Initiatives to Encourage More Hospitals to Migrate Their HISs to the Cloud

The primary objective of this strategy is to address the issue of "whether or not to adopt cloud technology." Based on the outcomes of the SWOT analysis, despite the pressing need for cloud technology to enhance health-centered medical service delivery today, hospitals remain cautious about its implementation due to external threats and internal weaknesses. Furthermore, providing cloud-based medical services has brought about a significant transformation within the medical industry that exceeds traditional independent operations and self-financing models used by hospitals. Therefore, governments should make greater efforts by implementing more active measures such as policy guidance, training planning, demonstration projects, or provision of cloud vouchers, in order to encourage more hospitals to securely migrate their HISs onto the cloud and meet demands for innovative medical services in this modern era.

Enhancing Hospitals' Overall Capability for Cloud Migration of HISs

The strategy aims to address the issue of "what preparations are necessary." As previously mentioned, cloud migration of HISs is a highly intricate system engineering project that requires hospitals to be fully prepared for its successful implementation. These preparations encompass various aspects including, but not limited to the following. First, human resources: hospitals should enhance employees' medical information literacy and application skills through comprehensive training programs, specialized lectures, or successful case studies. Second, material resources: hospitals should redesign and integrate existing systems and resources based on future development, existing foundations, and expert recommendations in order to optimize the use of cloud resources for acquiring more valuable information. Third, financial resources: hospitals require long-term financial investment planning to support the provision of cloud-based medical services. Moreover, health authorities should acknowledge that primary hospitals serve as both frontline institutions addressing medical needs and significant sources of authentic data. Therefore, moderate increases in investment in HIS construction of primary hospitals are necessary to ensure a continuous input of firsthand authentic data. Fourth, tools: a hospital that has robust IT capabilities can leverage free cloud migration consultation and tools provided by major cloud providers such as Alibaba, Tencent, Google, Microsoft, and Amazon Web Services, which can expedite the process of migrating information systems. However, it should be noted that the cloud services used (eg, computing and storage) may incur charges.

Establishing a Provincial-Level Unified Medical Hybrid Multi-Cloud Platform

The strategy aims to address the issue of "how to implement changes efficiently." In response to numerous complex internal and external situations, this study proposes a coping strategy: establishing a unified medical hybrid multi-cloud platform in each province or municipality.

First, the hybrid multi-cloud platform highly aligns with hospital operations. Hospitals require private clouds for storing sensitive and core data, nonpublic community clouds for internal consultations and other collaborations, public clouds for providing more extensive medical services to the public, and adopting a "multi-cloud" strategy to reduce risks such as vendor lock-in or declining service quality.

Second, a medical cloud platform at the provincial or municipal level can achieve maintainable security and more highly effective cloud migration. In comparison to hospitals, health authorities at this level possess stronger technological expertise, greater manpower resources, maintainable financial support, and relevant assets for constructing comprehensive platforms while effectively mitigating internal and external risks. Moreover, this approach can also foster overall advancements in cloud migration and system function quality of hospitals (particularly primary ones), as well as minimize redundant construction and maintenance expenses.

Last but most importantly, the scale of data possessed by one or several individual hospitals is insufficient to constitute true big data, limiting the opportunities for leveraging cloud computing's powerful intelligent analysis capabilities in uncovering hidden objective laws that can support the government to make scientific decisions. Considering factors such as data scale, cloud computing capabilities, and government economic support capacity, a provincial or municipal regional medical cloud platform is a more suitable choice.

Enhancing Legal Framework and Technical Support for Cloud-Based HISs

The primary goal of this strategy is to address the issue of "how to create a supportive environment." As an ancient Chinese proverb states, "A single log cannot support a crumbling building," indicating that relying solely on a provincial-level medical cloud platform is still insufficient to counter all external threats and internal risks faced by hospitals. Therefore, it requires a more proactive role from the government, which strengthens cooperation with relevant departments and enterprises to build a more robust and secure supporting environment for medical cloud platforms. Specifically, 2 aspects of support are necessary. First, credible legal support: although the Chinese government has been improving laws regarding cybersecurity, personal information protection, and data security, its support for cloud-based medical services remains inadequate. For example, in resolving disputes related to cloud medical services, health authorities still rely on laws such as the Physician Practice Law and Regulations on Prevention and Handling of Medical Disputes in China. However, these regulations have not explicitly defined the status and responsibilities of all parties involved in cloud-based medical services, which poses challenges in terms of judgments and penalties [26]. Therefore, it is essential to further refine relevant legislation and update existing regulations regarding doctors' practice rights, insurance responsibility, and reimbursement for medical insurance, ensuring doctors and patients can confidently participate in cloud-based medical services. Second, holistic technical support: as indicated by SWOT analysis results, hospitals often lack professionals with deep knowledge of cutting-edge technologies like cloud computing, AI, and big data. Therefore, establishing an organization like a medical cloud technology association becomes necessary. This organization should consist of experts from various relevant fields, including IT, communication, engineering, cryptography, medicine, and more. Their responsibilities would include devising unified long-term plans and action plans, standardizing contracts between hospitals and cloud service providers, reviewing hospitals' cloud migration plans and contracts, and conducting regular evaluations of the construction and operation of cloud-based HISs.

Conclusions

In conclusion, cloud computing is prioritized by the Chinese government as a "national strategic emerging industry." Despite encountering numerous challenges, the cloud migration of HISs in China exemplifies a prevailing development trend. Therefore, hospitals should adopt an open mindset and focus on enhancing their capabilities to develop medical services based on the cloud. Health authorities should also use more effective strategies to incentivize hospitals to migrate their HISs safely to the cloud, thereby fostering the flourishing growth of a novel health-centered medical industry.

The main contribution of this study is a comprehensive literature review and systematic SWOT analysis on the current status of cloud migration of HISs in China, and corresponding strategies for different combinations of internal and external influencing factors. It can help hospitals gain a clearer understanding of the overall situation while having more specific goals and methods when planning and implementing related work. Moreover, it can serve as a foundation for health authorities to develop policies aligned with the development of hospital informatization in the new era, as well as provide references for other countries or regions facing similar challenges.

There are 2 limitations in this study: first, not all personnel from hospitals contribute to writing papers, thus limiting the comprehensiveness of literature information; second, the SWOT analysis method assumes a distinction between internal and external factors, as well as a differentiation between strengths and weaknesses, overlooking the interrelated effects among influencing factors. To supplement and improve these aspects, more empirical investigation data are needed, along with a more rigorous analysis of the interactions among influencing factors. This will be the focus of my future research.

Data Availability

The data sets generated and analyzed during this study are not publicly available but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Academic paper database.
[XLSX File (Microsoft Excel File), 50 KB - medinform_v12i1e52080_app1.xlsx]

Multimedia Appendix 2
Supplementary database.
[XLSX File (Microsoft Excel File), 13 KB - medinform_v12i1e52080_app2.xlsx]

References

1. Zhou JY, Jiang Q, Ren J. The role and impact of cloud computing in hospital information management. Mod Hosp 2023 Mar 27;23(03):422-424 [FREE Full text] [doi: 10.3969/j.issn.1671-332X.2023.03.027]
2. Jia YW, Zhao D, Jiang KY, Luan GC. The U.S. federal government's cloud computing strategy. E-Government 2011 Jul 1(7):2-16 [FREE Full text] [doi: 10.16582/j.cnki.dzzw.2011.07.001]
3. Unleashing the potential of cloud computing in Europe. European Commission. 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52012DC0529 [accessed 2023-05-12]
4. Outline of the 14th five-year plan (2021-2025) for national economic and social development and vision 2035 of the People's Republic of China. The Government of China. 2021 Mar 12. URL: http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm [accessed 2023-05-12]
5. He XY, Wang LL, Wang L, Gao JH, Cui FF, Ma QQ, et al. Effectiveness of a cloud-based telepathology system in China: large-sample observational study. J Med Internet Res 2021;23(7):e23799 [FREE Full text] [doi: 10.2196/23799] [Medline: 34326037]
6. Gong MC, Liu L, Sun X, Yang Y, Wang S, Zhu H. Cloud-based system for effective surveillance and control of COVID-19: useful experiences from Hubei, China. J Med Internet Res 2020;22(4):e18948 [FREE Full text] [doi: 10.2196/18948] [Medline: 32287040]
7. Catteddu D, Hogben G. Cloud Computing Risk Assessment. The European Union Agency for Cybersecurity (ENISA). 2009 Nov. URL: https://www.enisa.europa.eu/publications/cloud-computing-risk-assessment [accessed 2023-05-16]
8. Wu J, Wang J, Nicholas S, Maitland E, Fan QY. Application of big data technology for COVID-19 prevention and control in China: lessons and recommendations. J Med Internet Res 2020;22(10):e21980 [FREE Full text] [doi: 10.2196/21980] [Medline: 33001836]
9. Healthcare cloud computing market. MarketsandMarkets. 2022. URL: https://www.marketsandmarkets.com/Market-Reports/cloud-computing-healthcare-market-347.html [accessed 2023-05-18]
10. Alipour J, Mehdipour Y, Karimi A, Sharifian R. Affecting factors of cloud computing adoption in public hospitals affiliated with Zahedan University of Medical Sciences: a cross-sectional study in the Southeast of Iran. Digit Health 2021;7:20552076211033428 [FREE Full text] [doi: 10.1177/20552076211033428] [Medline: 34777850]
11. Nikolopoulos M, Karampela I, Tzortzis E, Dalamaga M. Deploying cloud computing in the Greek healthcare system: a modern development proposal incorporating clinical and laboratory data. Stud Health Technol Inform 2018;251:35-38. [Medline: 29968595]
12. Rajendran J. What CFOs should know before venturing into the cloud. Healthc Financ Manage 2013;67(5):40-43. [Medline: 23678688]
13. Gao SM. Network security problems and countermeasures of hospital information system after going to the cloud. Comput Math Methods Med 2022;2022:9725741 [FREE Full text] [doi: 10.1155/2022/9725741] [Medline: 35898480]
14. Puyt R, Lie FB, De Graaf FJ, Wilderom CPM. Origins of SWOT analysis. Acad Manag Proc 2020;2020(1):17416. [doi: 10.5465/ambpp.2020.132]
15. The American Society of International Law. Congress enacts the Clarifying Lawful Overseas Use of Data (CLOUD) act, reshaping U.S. law governing cross-border access to data. Am J Int law 2018;112(3):487-493 [FREE Full text] [doi: 10.1017/ajil.2018.61]
16. Opara-Martins J, Sahandi R, Tian F. Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. J Cloud Comp 2016;5:4 [FREE Full text] [doi: 10.1186/s13677-016-0054-z]
17. Xiao Q. Research on Internet Medical Legal Regulation [Dissertation]. Shenzhen University. 2019. URL: https://cdmd.cnki.com.cn/Article/CDMD-10590-1019908748.htm [accessed 2023-10-18]
18. Berisha B, Mëziu E, Shabani I. Big data analytics in cloud computing: an overview. J Cloud Comput (Heidelb) 2022;11(1):24 [FREE Full text] [doi: 10.1186/s13677-022-00301-w] [Medline: 35966392]
19. Kong L. Application and exploration of collaborative medicine based on cloud computing. Dissertation. 2017. URL: https://d.wanfangdata.com.cn/thesis/Y3337261 [accessed 2023-08-25]

20. Zhang XG. Situation and strategy for the development of hospital informatization in the new era. China J Health Inform Manag 2018;15(4):367-372 [FREE Full text] [doi: 10.3969/j.issn.1672-5166.2018.04.01]
21. Setyonugroho W, Puspitarini AD, Kirana YC, Ardiansyah M. The complexity of the Hospital Information System (HIS) and obstacles in implementation: a mini-review. Enferm Clin 2020;30:233-235 [FREE Full text] [doi: 10.1016/j.enfcli.2020.06.053]
22. Zhang YX, Hu JP, Zhou GH, Xu XD. The development and prospect of health informatization during the 13th Five-Year Plan period. Chin J Health Inform Manag 2021 Jun 28;18(3):297-302 [FREE Full text] [doi: 10.1007/s35764-016-0106-7]
23. Zhao X, Li XH. Thoughts on the development of hospital information construction during the "14th five-year plan". Chin Hosp 2021;25(1):64-66 [FREE Full text] [doi: 10.19660/j.issn.1671-0592.2021.1.20]
24. Li HX, Xu F, Wang K. Research on the current situation of hospital informatization personnel configuration in China: a cross-sectional study. China Health Qual Manag 2022;01:4-7 [FREE Full text] [doi: 10.13912/j.cnki.chqm.2022.29.01.02]
25. Hao XN, Ma CY, Liu ZY, Zhou YC, Liu QK, Zhang S. The effects and problems on the reform of primary health informatization in China. Health Econ Res 2019;07:3-5 [FREE Full text]
26. Jiao YL. Investigation on legal status of internet hospital: From the perspective of "Ningbo cloud hospital". Chin J Health Pol 2017;10(10):72-75 [FREE Full text] [doi: 10.3969/j.issn.1674-2982.2017.10.012]
27. European digital infrastructure and data sovereignty-a policy perspective. European Institute of Innovation & Technology. 2020. URL: https://www.eitdigital.eu/fileadmin/files/2020/publications/data-sovereignty/EIT-Digital-Data-Sovereignty-Summary-Report.pdf [accessed 2023-05-18]
28. Decision on accelerating the cultivation and development of strategic emerging industries. The Government of China. 2010 Oct 18. URL: https://www.gov.cn/zwgk/2010-10/18/content_1724848.htm [accessed 2023-05-24]
29. Assessment report on the global computing index 2022-2023. Institute for Global Industry of Tsinghua University. 2023. URL: https://www.igi.tsinghua.edu.cn/info/1019/1321.htm [accessed 2023-11-16]
30. Cloud computing white paper (2023). China Academy of Information and Communications Technology (CAICT). 2023. URL: http://www.caict.ac.cn/kxyj/qwfb/bps/202307/P020230725521473129120.pdf [accessed 2023-11-16]
31. Zhang XX. The "2021 Internet Hospital Report" has been issued, incorporating analyses from 1,140 data sources and comprehensive examinations from 109 dimensions, thereby unearthing these core trends. Vbdata. 2021. URL: https://www.vbdata.cn/52404 [accessed 2023-06-01]
32. Hospital cloud service application survey report. China Hospital Information Management Association (CHIMA). 2022. URL: https://www.hit180.com/57127.html [accessed 2023-06-23]
33. Bulletin of the seventh national population census (No. 5) - age composition of the population. National Bureau of Statistics of China (CNBS). 2021. URL: http://www.stats.gov.cn/sj/tjgb/rkpcgb/qgrkpcgb/202302/t20230206_1902005.html [accessed 2023-06-02]
34. Guiding opinions on establishing and improving the healthcare system for the elderly. National Health Commission of China (CNHC). 2019. URL: http://www.nhc.gov.cn/lljks/s7785/201911/cf0ad12cb0ec4c96b87704fbbeb5bbde.shtml [accessed 2023-06-16]
35. The cloud 100. Forbes. 2023. URL: https://www.forbes.com/lists/cloud100/?sh=6d17eead7d9c [accessed 2023-09-16]
36. Report on trade practices in cloud services sector. Japan Fair Trade Commission (JFTC). 2022. URL: https://www.jftc.go.jp/en/pressreleases/yearly-2022/June/220722_2EN.pdf [accessed 2023-11-16]
37. Hardesty L. European cloud providers take hit from AWS, Google, Azure, says Synergy. Silverlinings. 2021 Sep 23. URL: https://www.silverliningsinfo.com/platforms/european-cloud-providers-take-hit-from-aws-google-azure-says-synergy [accessed 2023-12-30]
38. Huawei Cloud won the bid of the Shijiazhuang Beiguo Electronics public cloud project: Ali Cloud, Unicom, Telecom, and Xinhua Net lost the bid. NetEase. 2021. URL: https://www.163.com/dy/article/GPRKUNKM05386WWT.html [accessed 2023-10-13]
39. Li RZZ, Hua J. The global healthcare sector often faced ransomware attacks; health authorities should prioritize cybersecurity preparedness. SFC. URL: https://www.sfccn.com/2022/2-10/2MMDE0MDVfMTY5NjM2Mw.html [accessed 2023-10-03]
40. Nipitpon SA. Why the cloud is a lifeline for the NHS. Open Access Government. 2020. URL: https://www.openaccessgovernment.org/why-the-cloud-is-a-lifeline-for-the-nhs/84112/ [accessed 2023-09-21]
41. Realising the value of health care data: a framework for the future. Ernst & Young Global Limited. 2020. URL: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/life-sciences/life-sciences-pdfs/ey-value-of-health-care-data-v20-final.pdf [accessed 2023-10-12]
42. Xu L, Wu DP, Li Y, Zhang L, Wang Y, Wang QY, et al. Application of portable electroencephalograph in patients with epilepsy and establishment of medical platform. National Medical Journal of China 2020;100(47):3764-3767. [doi: 10.3760/cma.j.cn112137-20200703-02023] [Medline: 33379840]
43. Survey report on the informationization status of Chinese hospitals in 2019-2020 (public version). China Hospital Information Management Association (CHIMA). 2021. URL: https://www.chima.org.cn/Html/News/Articles/8684.html [accessed 2023-10-12]
44. Xie XX, Zhou WM, Lin LY, Fan S, Lin F, Wang L, et al. Internet hospitals in China: cross-sectional survey. J Med Internet Res 2017;19(7):e239 [FREE Full text] [doi: 10.2196/jmir.7854] [Medline: 28676472]

45. Liu D, Li T, Liu X, Wang DF. Survey and analysis on digital construction of primary hospitals in Guizhou Province. Chinese Critical Care Medicine 2022 Aug;34(8):863-870. [doi: 10.3760/cma.j.cn121430-20220511-00476] [Medline: 36177932]

Abbreviations

AI: artificial intelligence
CAGR: compound annual growth rate
CLOUD Act: Clarifying Lawful Use of Data Abroad Act
CNCTPS: Chinese National Cloud-Based Telepathology System
DOI: digital object identifier
HIS: hospital information system
NHS: National Health Service
PMID: PubMed unique identifier
SWOT: Strengths, Weaknesses, Opportunities, and Threats

Edited by C Lovis; submitted 22.08.23; peer-reviewed by T Khodaveisi, M Platt, C Xie; comments to author 07.10.23; revised version received 30.11.23; accepted 02.12.23; published 05.02.24.

Please cite as:
Xu J
The Current Status and Promotional Strategies for Cloud Migration of Hospital Information Systems in China: Strengths, Weaknesses, Opportunities, and Threats Analysis
JMIR Med Inform 2024;12:e52080
URL: https://medinform.jmir.org/2024/1/e52080
doi:10.2196/52080
PMID:38315519

Viewpoint

A Roadmap for Using Causal Inference and Machine Learning to Personalize Asthma Medication Selection

Flory L Nkoy[1*], MS, MPH, MD; Bryan L Stone[1], MS, MD; Yue Zhang[2,3], PhD; Gang Luo[4*], PhD

[1]Department of Pediatrics, University of Utah, Salt Lake City, UT, United States

[2]Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City, UT, United States

[3]Division of Biostatistics, Department of Population Health Sciences, University of Utah, Salt Lake City, UT, United States

[4]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

[*]these authors contributed equally

Corresponding Author:
Gang Luo, PhD
Department of Biomedical Informatics and Medical Education
University of Washington
850 Republican Street, Building C
Box 358047
Seattle, WA, 98195
United States
Phone: 1 2062214596
Fax: 1 2062212671
Email: gangluo@cs.wisc.edu

Abstract

Inhaled corticosteroid (ICS) is a mainstay treatment for controlling asthma and preventing exacerbations in patients with persistent asthma. Many types of ICS drugs are used, either alone or in combination with other controller medications. Despite the widespread use of ICSs, asthma control remains suboptimal in many people with asthma. Suboptimal control leads to recurrent exacerbations, causes frequent ER visits and inpatient stays, and is due to multiple factors. One such factor is the inappropriate ICS choice for the patient. While many interventions targeting other factors exist, less attention is given to inappropriate ICS choice. Asthma is a heterogeneous disease with variable underlying inflammations and biomarkers. Up to 50% of people with asthma exhibit some degree of resistance or insensitivity to certain ICSs due to genetic variations in ICS metabolizing enzymes, leading to variable responses to ICSs. Yet, ICS choice, especially in the primary care setting, is often not tailored to the patient's characteristics. Instead, ICS choice is largely by trial and error and often dictated by insurance reimbursement, organizational prescribing policies, or cost, leading to a one-size-fits-all approach with many patients not achieving optimal control. There is a pressing need for a decision support tool that can predict an effective ICS at the point of care and guide providers to select the ICS that will most likely and quickly ease patient symptoms and improve asthma control. To date, no such tool exists. Predicting which patient will respond well to which ICS is the first step toward developing such a tool. However, no study has predicted ICS response, forming a gap. While the biologic heterogeneity of asthma is vast, few, if any, biomarkers and genotypes can be used to systematically profile all patients with asthma and predict ICS response. As endotyping or genotyping all patients is infeasible, readily available electronic health record data collected during clinical care offer a low-cost, reliable, and more holistic way to profile all patients. In this paper, we point out the need for developing a decision support tool to guide ICS selection and the gap in fulfilling the need. Then we outline an approach to close this gap via creating a machine learning model and applying causal inference to predict a patient's ICS response in the next year based on the patient's characteristics. The model uses electronic health record data to characterize all patients and extract patterns that could mirror endotype or genotype. This paper supplies a roadmap for future research, with the eventual goal of shifting asthma care from one-size-fits-all to personalized care, improve outcomes, and save health care resources.

(JMIR Med Inform 2024;12:e56572) doi:10.2196/56572

KEYWORDS

asthma; causal inference; forecasting; machine learning; decision support; drug; drugs; pharmacy; pharmacies; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutics; pharmaceuticals; pharmaceutical; medication; medications; medication selection;

respiratory; pulmonary; forecast; ICS; inhaled corticosteroid; inhaler; inhaled; corticosteroid; corticosteroids; artificial intelligence; personalized; customized

Introduction

Asthma is a chronic disease characterized by inflammation, narrowing, and hyperactivity of the airways causing shortness of breath, chest tightness, coughing, and wheezing [1]. Asthma affects about 25 million people in the United States [2]. In 2021, there were 9.8 million exacerbations of asthma symptoms (or asthma attacks) leading to over 980,000 emergency room (ER) visits and over 94,500 hospitalizations [2]. Asthma costs the US economy over US $80 billion in health care expenses each year, work and school absenteeism, and deaths [3].

Inhaled corticosteroid (ICS) is a mainstay treatment for controlling asthma and preventing exacerbations in patients with persistent asthma [4] accounting for over 60% of people with asthma [5,6]. Many types of ICS drugs are used, either alone like fluticasone (Flovent, Arnuity, and Aller-flo), budesonide (Pulmicort, Entocort, and Rhinocort), mometasone (Asmanex), beclomethasone (Beclovent, Qvar, Vancenase, Beconase, Vanceril, and Qnasl), ciclesonide (Alvesco), and so forth, or in combination with a long-acting beta2 agonist like fluticasone/salmeterol (Advair), budesonide/formoterol (Symbicort), mometasone/formoterol (Dulera), and fluticasone/vilanterol (Breo), and so forth [4]. Regular use of appropriate ICSs improves asthma control and reduces airway inflammation, symptoms, exacerbations, ER visits, and inpatient stays [7-9].

Despite the widespread use of ICSs, asthma control remains suboptimal in many people with asthma [10-13] including 44% of children and 60% of adults based on asthma exacerbations in the past year [14,15], 72% of patients based on asthma control test [10], 53% of children and 44% of adults based on asthma attacks in the past year [16], and 59% of children based on the 2007-2013 Medical Expenditure Panel Survey [17]. Suboptimal control leads to recurrent exacerbations, causes frequent ER visits and inpatient stays, and is projected to have an economic burden of US $963.5 billion over the next 20 years [18]. Suboptimal control is due to multiple factors [19-23] including (1) failure to recognize and act on early signs of declining control [24,25], (2) lack of self-management skills, (3) nonadherence to therapy [26], and (4) inappropriate ICS choice for the patient [27-32]. While interventions targeting other factors exist, less attention has been given to inappropriate ICS choice.

Asthma is heterogeneous with variable profiles in terms of clinical presentations (phenotypes) and underlying mechanisms (endotypes) [33,34]. Molecular techniques have revealed a few phenotype and endotype relationships, allowing the categorization of asthma into two main groups (1) T-helper type 2 (Th2)-high (eg, atopic and late onset) and (2) Th2-low (eg, nonatopic, smoking-related, and obesity-related) [33,34]. It is known that within the 2 groups, there are many subgroups [33,35] with different biomarker expressions (eg, immunoglobulin E [IgE], fractional exhaled nitric oxide [FeNO], interleukin [IL]-4, IL-5, and IL-13) [36]. So far, only a few biomarkers have been characterized for use in clinical practice. Despite a few successes using biomarkers for targeted therapy, ICS choice, especially in the primary care setting, is largely by trial and error and many patients remain uncontrolled [37-42].

Besides patient nonadherence and environmental factors, response to ICS treatment is affected by genetic variations in ICS metabolizing enzymes [43,44], regardless of whether the ICS is used alone or is combined with another asthma medication like a long-acting beta2 agonist. Single nucleotide polymorphisms in cap methyltransferase 1 (CMTR1), tripartite motif containing 24 (TRIM24), and membrane associated guanylate kinase, WW and PDZ domain containing 2 (MAGI2) genes were found to be associated with variability in asthma exacerbations [43]. Additional evidence supports that these genes also cause variability in ICS response [44]. Due to genetic variations in cytochrome P (CYP) 450 enzymes that metabolize over 80% of drugs including ICS, up to 50% of people with asthma have altered metabolism to certain ICSs [45-51] impacting asthma control [52,53]. CYP3A5*3/*3 and CYP3A4*22 genotypes were found to be linked to ICS response [54,55]. These studies provide evidence that genetic variations greatly affect ICS responsiveness, although the exact relationships between genetic variations and ICS response remain largely unknown [36,56,57]. Currently, many candidate genes are being studied, and pharmacogenetics has not yet reached routine clinical practice in asthma care.

ICS choice for patients is often dictated by insurance reimbursement, organizational policies, or cost, leading to a one-size-fits-all approach [37-42]. Some insurers require patients to first fail on a cheaper ICS before authorizing a more expensive ICS [39]. Nonmedical switch due to preferred drug formulary change is common and leads to bad outcomes, with 70% of patients reporting more exacerbations after the switch [39]. Patients also often report that they tried a few different ICSs before ending up with the drug that gave them the most relief, with 60% reporting it was hard for their providers to find the effective drug [37-39]. Cycling through various ICSs delays the start of an effective ICS and is neither efficient nor cost-effective [39]. New strategies are needed to allow a faster and more efficient way to tailor ICS selection to each patient's characteristics [36].

While the biologic heterogeneity of asthma is vast, few, if any, biomarkers or genotypes can currently be used to systematically profile all patients with asthma and predict ICS response [36,58,59]. Readily available electronic health record (EHR) data collected during clinical care offer a low-cost, reliable, and more holistic way to profile all patients [36,60]. With a high accuracy of 87%-95% [36], machine learning models using EHR data have been used to profile patients in various areas, for example, to develop a phenotype for patients with Turner syndrome [61], identify low medication adherence profiles [62], find variable COVID-19 treatment response profiles [63], and predict hypertension treatment response [64]. Yet, while machine learning has helped find various asthma profiles [65-72], no prior study has predicted ICS response. Also, prior

studies are mostly from single centers with small sample sizes and have not moved the needle of precision treatment for asthma [58,60].

A decision support tool is greatly needed, especially in the primary care setting, to guide providers to select at the point of care the ICS that will most likely and quickly ease patient symptoms and improve asthma control. Forecasting which patient will respond well to which ICS is the first step toward creating this tool, but no prior study has predicted ICS response, forming a gap.

To shift asthma care from one-size-fits-all to personalized care, improve outcomes, and save health care resources, we make three contributions in this paper, supplying a roadmap for future research: (1) we point out the above-mentioned need for creating a decision support tool to guide ICS selection; (2) we point out the above-mentioned gap in fulfilling this need; and (3) to close this gap, we outline an approach to create a machine learning model and apply causal inference to predict a patient's ICS response in the next year based on the patient's characteristics. We present the central ideas of this approach in the following sections.

Creating a Machine Learning Model and Applying Causal Inference to Predict ICS Response

Overview of Our Approach

We use EHR data from a large health care system to develop a machine learning model and apply casual inference to predict a patient's ICS response based on the patient's characteristics. As endotyping or genotyping all patients is infeasible, our model uses EHR data to characterize all patients and extract patterns that could mirror endotype or genotype. Our model is trained on historical data, and can then be applied to new patients to guide ICS selection during an initial or early encounter for asthma care. The optimal ICS choice identified by our approach can be either an ICS (generic name and dosage) alone or an ICS combined with another asthma medication like a long-acting beta2 agonist.

Both pediatric and adult patients with asthma are treated by primary care providers (PCPs) who are mostly generalists and asthma specialists including allergists, immunologists, and pulmonologists. Large differences exist between PCPs and specialists in terms of knowledge, care patterns, and asthma outcomes, with asthma specialists adhering more often to guideline recommendations [73-76]. A greater difference exists between PCPs and specialists in controller medication use [76]. Compared to PCPs, asthma specialists tend to achieve better outcomes [77], including higher physical functioning [78], better patient-reported care [78], and fewer ER visits and inpatient stays [78-84]. As over 60% of people with asthma are cared for by PCPs [85], our machine learning model primarily targets PCPs, although asthma specialists could also benefit from this model.

The asthma medication ratio (AMR) is the total number of units of asthma controller medications dispensed divided by the total number of units of asthma medications (controllers + relievers) dispensed [86,87]. Higher AMR (≥0.5) is associated with less oral corticosteroid use (a surrogate measure for asthma exacerbations), fewer ER visits and inpatient stays, and lower costs [87-89]. Lower AMR (<0.5) is associated with more exacerbations, ER visits, and inpatient stays [90,91]. Approved by Healthcare Effectiveness Data and Information Set (HEDIS) as a quality measure, AMR is widely used by health care systems [89]. AMR is a reliable reflection of asthma control and gives an accurate assessment of asthma exacerbation risk [92]. We use change in AMR as the prediction target of our model for predicting ICS response, as AMR can be calculated on all patients. In comparison, neither asthma control nor acute outcomes (eg, ER visits, inpatient stays, or oral corticosteroid use) is used as the prediction target, as the former is often missing in EHRs and the latter does not occur in all patients. An effective ICS will lead to less reliever use and increased AMR. An ineffective ICS will lead to more reliever use and reduced AMR. We formerly used EHR data to build accurate models to predict hospital use (ER visit or inpatient stay) for asthma [93-95]. We expect EHR data to have great predictive power for AMR, which is associated with hospital use for asthma [87-91]. Using the AMR can facilitate the dissemination of our approach across health care systems.

We outline the individual steps of our approach in the following sections.

Step 1: Building a Machine Learning Model to Predict a Patient's ICS Response Defined by Changes in AMR

We focus on patients with persistent asthma for whom ICSs are mainly used. We use the HEDIS case definition of persistent asthma [96,97], the already validated [98] and the most commonly used administrative data marker of persistent asthma [97]. A patient is deemed to have persistent asthma if in each of 2 consecutive years, the patient meets at least one of the following criteria: (1) at least 1 ER visit or inpatient stay with a principal diagnosis code of asthma (*ICD-9* [*International Classification of Diseases, Ninth Revision*] 493.0x, 493.1x, 493.8x, 493.9x; *ICD-10* [*International Classification of Diseases, Tenth Revision*] J45.x), (2) at least 2 asthma medication dispensing and at least 4 outpatient visits, each with a diagnosis code of asthma, and (3) at least 4 asthma medication dispensing. In the rest of this paper, we always use patients with asthma to refer to patients with persistent asthma. The prediction target or outcome is the amount of change in a patient's AMR after 1 year. The AMR is computed over a 1-year period [86,87].

We combine patient, air quality, and weather features computed on the raw variables to build the model to predict ICS response. Existing predictive models for asthma outcomes [93-95,99-110] rarely use air quality and weather variables, but these variables impact asthma outcomes [111-117] (eg, short-term exposure to air pollution, even if measured at the regional level, is associated with asthma exacerbations [113-117]). For each such variable, we examine multiple features (eg, mean, maximum, SD, and slope). We examine over 200 patient features listed in our papers' [93-95] appendices and formerly used to predict hospital use for asthma, which is associated with AMR [87-91]. Several examples of these features are comorbidities, allergies, the

number of the patient's asthma-related ER visits in the prior 12 months, the total number of units of systemic corticosteroids ordered for the patient in the prior 12 months, and the number of primary or principal asthma diagnoses of the patient in the prior 12 months. We also use as features the patient's current AMR computed over the prior 12 months [86,87], the generic name and the dosage of the ICS that the patient currently uses, and those of the long-acting beta2 agonist, leukotriene receptor antagonist, biologic or another asthma medication, if any, that is combined with the ICS.

Step 2: Conducting Causal Machine Learning to Identify Optimal ICS Choice

Our goal is to integrate machine learning and G-computation to develop a method to estimate the causal effects of various ICS choices on AMR for patients with specific characteristics. This causal machine learning method [118] processes large data sets by capturing complex nonlinear relationships between features, thereby revealing the cause-and-effect relationships between ICS choice and change in AMR. We use the machine learning model built in step 1. Using G-computation [119,120], an imputation-based causal inference method, we estimate the potential effects of hypothetical ICS choices with specific dosages on changes in AMR after 1 year. G-computation builds on the machine learning model of the outcome as a function of ICS indicators, ICS dosages, and other features to predict AMR outcomes under different counterfactual ICS choice scenarios. CIs are estimated through 10,000 bootstrap resampling with replacement [121].

We apply causal machine learning to estimate the impact of ICS choices on patients with specific characteristics by averaging predicted AMR after 1 year for a given ICS and these characteristics across all participants. This estimation is contrasted with the averaged predicted outcome in the absence of any ICS choice. The ICS choice with the highest and statistically significant contrast estimation is identified as the optimal choice for patients with these characteristics. All hypotheses can be tested at a significance level of .05.

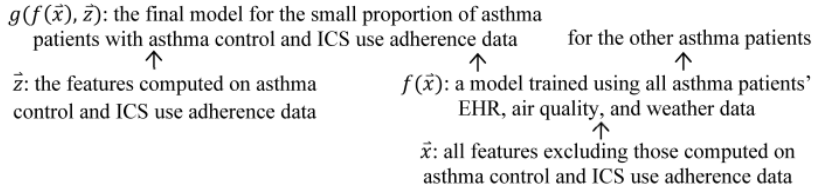
Step 3: Assessing the Impact of Adding External Patient-Reported Asthma Control and ICS Use Adherence Data on the Model's Predictions

EHRs have limitations regarding patient-reported data with extra predictive power such as asthma control and ICS use adherence. For asthma, asthma control and ICS use adherence are critical variables, as (1) a patient's asthma control fluctuates over time and drives the provider's decision to prescribe or adjust ICSs and (2) ICS use adherence impacts the patient's asthma control and helps assess whether the patient is actually responding to an ICS. However, despite their high predictive power for patient outcomes, these variables are not routinely collected or included in EHRs in clinical practice. At Intermountain Healthcare, the largest health care system in Utah, we pioneered the electronic AsthmaTracker, a mobile health (mHealth) app used weekly to assess, collect, and monitor patients' asthma control and actual ICS use adherence [122]. Like most patient-reported data, these patient-reported variables have been collected on only a small proportion of patients with asthma. To date, 1380 patients with asthma have used the app and produced about 45,000 records of weekly asthma control scores and ICS use adherence data (eg, the ICS' name and the number of days an ICS is actually used by the patient in that week). If we train a predictive model using EHR and patient-reported data limited to this small proportion of patients, the model will be inaccurate due to insufficient training data. Yet, for these patients, combining their patient-reported data with the outputs of a model built on all patients' EHR data can help raise the prediction accuracy for them. To realize this, we propose the first method to combine external patient-reported data available on a small proportion of patients with the outputs of a model built on all patients' EHR data to raise prediction accuracy for the small proportion of patients while maintaining prediction accuracy for the other patients.

To illustrate how our method works, we consider the case that the model created in step 1 is built using Intermountain Healthcare EHR data. The weekly asthma control scores and ICS use adherence data collected from the 1380 patients with asthma are unused in step 1. Now we add features (eg, mean, SD, and slope) computed on patient-reported asthma control and ICS use adherence data to raise prediction accuracy for these patients. Among all patients with asthma, only 1% have asthma control and ICS use adherence data. We use the method shown in Figure 1 to combine the asthma control and ICS use adherence data from this small proportion of patients with the outputs of a model trained on EHR, air quality, and weather data of all patients with asthma. We start from the original model built in step 1. This model is reasonably accurate, as it is trained using EHR, air quality, and weather data of all patients with asthma and all features excluding those computed on asthma control and ICS use adherence data. For each patient with asthma control and ICS use adherence data, we apply the model to the patient, obtain a prediction result, and use this result as a feature. We then combine this new feature with the features computed on asthma control and ICS use adherence data to train a second model for these patients using their data. The second model is built upon and thus tends to be more accurate than the original model for these patients. The original model is used for the other patients. Our method is general, works for all kinds of features, and is not limited to any specific disease, prediction target, cohort, or health care system. Whenever a small proportion of patients have extra predictive variables, we could use this method to raise prediction accuracy for these patients while maintaining prediction accuracy for the other patients.

For the patients with asthma control and ICS use adherence data, we compare the mean squared and the mean absolute prediction errors gained by the model built in step 1 and the second model built here. We expect adding asthma control and ICS use adherence data to the model to lower both prediction errors. The error drop rates help reveal the value of routinely collecting asthma control and ICS use adherence data in clinical care to lower prediction errors. Currently, such data are rarely collected.

Figure 1. Our method to raise prediction accuracy for the small proportion of patients with asthma and asthma control and ICS use adherence data while maintaining prediction accuracy for the other patients with asthma. EHR: electronic health record; ICS: inhaled corticosteroid.



Discussion

Principal Findings

Besides the variables mentioned in the "Step 1: Building a machine learning model to predict a patient's ICS response defined by changes in AMR" section, environmental variables beyond air quality and weather and many other factors can impact patient outcomes. Moreover, there are almost infinite possible features. For any first future study that one will do along the direction pointed out in this paper, a realistic goal is to show that using our methods can build decent models and improve asthma care rather than to exhaust all possible useful variables and features and obtain the theoretically highest possible model performance. Not accounting for all possible factors limits the generalizability of these models to medication selection for other diseases.

We use the G-computation method to conduct causal inference. This method relies heavily on correctly specifying the predictive model for ICS response, including accurately identifying all relevant confounders and interactions and incorporating them into the model. Misspecification of the model can lead to biased estimated effects of various ICS choices on AMR. To address this issue, we can adopt several preventive strategies during model development. We engage with subject matter experts to ensure that the model includes all relevant variables and reflects the underlying process. To guide model development and help identify potential sources of bias, we construct a directed acyclic graph that lays out the relationships among the independent and dependent variables. We use machine learning techniques that provide flexible modeling approaches to capture complex relationships among variables. When reporting our findings, we keep transparent about the final model specification and the rationale behind our model building process. We believe using these strategies will mitigate the risk of model misspecification and strengthen the reliability of our estimated effects of various ICS choices on AMR.

AMR is reported to be a reliable reflection of asthma control and of asthma exacerbation risk [92]. In a future study that we plan to do along the direction pointed out in this paper, we can use Intermountain Healthcare data to validate this relationship. Specifically, we use multivariable linear regression to assess the relationship between the AMR computed on EHR data and the patient's asthma control level obtained from the external patient-reported data, while controlling for other factors. We expect to see a strong and positive association between the AMR and the patient's asthma control level.

When creating the model in step 1, we can include medication persistence measures computed on insurance claim data [123], such as the proportion of days covered for ICS, as features. However, this does not obviate the need to examine patient-reported ICS use adherence data in step 3. ICS persistence measures give information on the possession of ICS, but not on actual use of ICS. Each ICS persistence measure is computed at a coarse time granularity as an average value over a long period. In comparison, our patient-reported ICS use adherence data offer information on the actual use of ICS. The data are at a fine time granularity, with 1 set of values per week for a patient. This enables us to compute features on various patterns and trends that can be useful for making predictions.

Conclusions

In asthma care, ICS choice is largely by trial and error and often made by a one-size-fits-all approach with many patients not achieving optimal outcomes. In this paper, we point out the need for creating a decision support tool to guide ICS selection and a gap in fulfilling this need. Then we outline an approach to close this gap via creating a machine learning model and applying causal inference to predict a patient's ICS response in the next year based on the patient's characteristics. This supplies a roadmap for future research.

Authors' Contributions

FLN and GL are co-senior authors mainly responsible for the paper. They conceptualized the presentation approach, performed literature review, and wrote the paper. BLS provided feedback on various medical issues, contributed to conceptualizing the presentation, and revised the paper. YZ wrote the causal inference section. All authors read and approved the final paper.

Conflicts of Interest

GL is an editorial board member of *JMIR AI*. The other authors declare no conflicts of interest.

References

1. Hashmi MF, Tariq M, Cataletto ME. Asthma. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.

2. Most recent national asthma data. Centers for Disease Control and Prevention. 2023. URL: https://www.cdc.gov/asthma/most_recent_national_asthma_data.htm [accessed 2024-01-22]
3. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc 2018;15(3):348-356 [FREE Full text] [doi: 10.1513/AnnalsATS.201703-259OC] [Medline: 29323930]
4. Inhaled corticosteroids. American Academy of Allergy, Asthma & Immunology. 2023. URL: https://www.aaaai.org/tools-for-the-public/drug-guide/inhaled-corticosteroids [accessed 2024-01-22]
5. Asthma severity among children with current asthma. Centers for Disease Control and Prevention. 2023. URL: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/asthma/asthma_stats/severity_child.htm [accessed 2024-01-22]
6. Asthma severity among adults with current asthma. Centers for Disease Control and Prevention. 2023. URL: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/asthma/asthma_stats/severity_adult.htm [accessed 2024-01-22]
7. Averell CM, Laliberté F, Germain G, Duh MS, Rousculp MD, MacKnight SD, et al. Impact of adherence to treatment with inhaled corticosteroids/long-acting β-agonists on asthma outcomes in the United States. Ther Adv Respir Dis 2022;16:17534666221116997 [FREE Full text] [doi: 10.1177/17534666221116997] [Medline: 36036456]
8. Cardet JC, Papi A, Reddel HK. "As-needed" inhaled corticosteroids for patients with asthma. J Allergy Clin Immunol Pract 2023;11(3):726-734 [FREE Full text] [doi: 10.1016/j.jaip.2023.01.010] [Medline: 36702246]
9. Sadatsafavi M, Lynd LD, De Vera MA, Zafari Z, FitzGerald JM. One-year outcomes of inhaled controller therapies added to systemic corticosteroids after asthma-related hospital discharge. Respir Med 2015;109(3):320-328 [FREE Full text] [doi: 10.1016/j.rmed.2014.12.014] [Medline: 25596136]
10. George M, Balantac Z, Gillette C, Farooqui N, Tervonen T, Thomas C, et al. Suboptimal control of asthma among diverse patients: a US mixed methods focus group study. J Asthma Allergy 2022;15:1511-1526 [FREE Full text] [doi: 10.2147/JAA.S377760] [Medline: 36313858]
11. Sullivan PW, Ghushchyan V, Kavati A, Navaratnam P, Friedman HS, Ortiz B. Trends in asthma control, treatment, health care utilization, and expenditures among children in the United States by place of residence: 2003-2014. J Allergy Clin Immunol Pract 2019;7(6):1835-1842.e2. [doi: 10.1016/j.jaip.2019.01.055] [Medline: 30772478]
12. Zhang S, White J, Hunter AG, Hinds D, Fowler A, Gardiner F, et al. Suboptimally controlled asthma in patients treated with inhaled ICS/LABA: prevalence, risk factors, and outcomes. NPJ Prim Care Respir Med 2023;33(1):19 [FREE Full text] [doi: 10.1038/s41533-023-00336-9] [Medline: 37156824]
13. Nurmagambetov TA, Krishnan JA. What will uncontrolled asthma cost in the United States? Am J Respir Crit Care Med 2019;200(9):1077-1078 [FREE Full text] [doi: 10.1164/rccm.201906-1177ED] [Medline: 31251082]
14. Uncontrolled asthma among children with current asthma, 2018-2020. Centers for Disease Control and Prevention. 2021. URL: https://tinyurl.com/ycdz2mp2 [accessed 2024-01-22]
15. Uncontrolled asthma among adults, 2019. Centers for Disease Control and Prevention. 2020. URL: https://archive.cdc.gov/#/details?url=https://www.cdc.gov/asthma/asthma_stats/uncontrolled-asthma-adults-2019.htm [accessed 2024-01-22]
16. Pate CA, Zahran HS, Qin X, Johnson C, Hummelman E, Malilay J. Asthma surveillance—United States, 2006-2018. MMWR Surveill Summ 2021;70(5):1-32 [FREE Full text] [doi: 10.15585/mmwr.ss7005a1] [Medline: 34529643]
17. Sullivan PW, Ghushchyan V, Navaratnam P, Friedman HS, Kavati A, Ortiz B, et al. National prevalence of poor asthma control and associated outcomes among school-aged children in the United States. J Allergy Clin Immunol Pract 2018;6(2):536-544.e1. [doi: 10.1016/j.jaip.2017.06.039] [Medline: 28847656]
18. Yaghoubi M, Adibi A, Safari A, FitzGerald JM, Sadatsafavi M. The projected economic and health burden of uncontrolled asthma in the United States. Am J Respir Crit Care Med 2019;200(9):1102-1112 [FREE Full text] [doi: 10.1164/rccm.201901-0016OC] [Medline: 31166782]
19. Centers for Disease Control and Prevention (CDC). Asthma hospitalizations and readmissions among children and young adults--Wisconsin, 1991-1995. MMWR Morb Mortal Wkly Rep 1997;46(31):726-729 [FREE Full text] [Medline: 9262074]
20. Li D, German D, Lulla S, Thomas RG, Wilson SR. Prospective study of hospitalization for asthma. A preliminary risk factor model. Am J Respir Crit Care Med 1995;151(3 Pt 1):647-655. [doi: 10.1164/ajrccm.151.3.7881651] [Medline: 7881651]
21. Crane J, Pearce N, Burgess C, Woodman K, Robson B, Beasley R. Markers of risk of asthma death or readmission in the 12 months following a hospital admission for asthma. Int J Epidemiol 1992;21(4):737-744. [doi: 10.1093/ije/21.4.737] [Medline: 1521979]
22. Mitchell EA, Bland JM, Thompson JM. Risk factors for readmission to hospital for asthma in childhood. Thorax 1994;49(1):33-36 [FREE Full text] [doi: 10.1136/thx.49.1.33] [Medline: 8153938]
23. Vargas PA, Perry TT, Robles E, Jo CH, Simpson PM, Magee JM, et al. Relationship of body mass index with asthma indicators in head start children. Ann Allergy Asthma Immunol 2007;99(1):22-28. [doi: 10.1016/S1081-1206(10)60616-3] [Medline: 17650825]
24. Barnes PJ. Achieving asthma control. Curr Med Res Opin 2005;21(Suppl 4):S5-S9. [doi: 10.1185/030079905X61730] [Medline: 16138939]
25. Bloomberg GR, Banister C, Sterkel R, Epstein J, Bruns J, Swerczek L, et al. Socioeconomic, family, and pediatric practice factors that affect level of asthma control. Pediatrics 2009;123(3):829-835 [FREE Full text] [doi: 10.1542/peds.2008-0504] [Medline: 19255010]

26. Bateman ED, Frith LF, Braunstein GL. Achieving guideline-based asthma control: does the patient benefit? Eur Respir J 2002;20(3):588-595 [FREE Full text] [doi: 10.1183/09031936.02.00294702] [Medline: 12358333]
27. Chapman KR, Boulet LP, Rea RM, Franssen E. Suboptimal asthma control: prevalence, detection and consequences in general practice. Eur Respir J 2008;31(2):320-325 [FREE Full text] [doi: 10.1183/09031936.00039707] [Medline: 17959642]
28. Rabe KF, Adachi M, Lai CK, Soriano JB, Vermeire PA, Weiss KB, et al. Worldwide severity and control of asthma in children and adults: the global asthma insights and reality surveys. J Allergy Clin Immunol 2004;114(1):40-47 [FREE Full text] [doi: 10.1016/j.jaci.2004.04.042] [Medline: 15241342]
29. National Asthma Education and Prevention Program. Expert Panel Report 3 (EPR-3): guidelines for the diagnosis and management of asthma-summary report 2007. J Allergy Clin Immunol 2007;120(Suppl 5):S94-S138. [doi: 10.1016/j.jaci.2007.09.043] [Medline: 17983880]
30. Stempel DA, McLaughin TP, Stanford RH, Fuhlbrigge AL. Patterns of asthma control: a 3-year analysis of patient claims. J Allergy Clin Immunol 2005;115(5):935-939 [FREE Full text] [doi: 10.1016/j.jaci.2005.01.054] [Medline: 15867848]
31. Cukovic L, Sutherland E, Sein S, Fuentes D, Fatima H, Oshana A, et al. An evaluation of outpatient pediatric asthma prescribing patterns in the United States. Int J Sci Res Arch 2023;9(1):344-349 [FREE Full text] [doi: 10.30574/ijsra.2023.9.1.0388]
32. Belhassen M, Nibber A, Van Ganse E, Ryan D, Langlois C, Appiagyei F, et al. Inappropriate asthma therapy-a tale of two countries: a parallel population-based cohort study. NPJ Prim Care Respir Med 2016;26:16076 [FREE Full text] [doi: 10.1038/npjpcrm.2016.76] [Medline: 27735927]
33. McIntyre AP, Viswanathan RK. Phenotypes and endotypes in asthma. Adv Exp Med Biol 2023;1426:119-142. [doi: 10.1007/978-3-031-32259-4_6] [Medline: 37464119]
34. Kuruvilla ME, Lee FE, Lee GB. Understanding asthma phenotypes, endotypes, and mechanisms of disease. Clin Rev Allergy Immunol 2019;56(2):219-233 [FREE Full text] [doi: 10.1007/s12016-018-8712-1] [Medline: 30206782]
35. Salter B, Lacy P, Mukherjee M. Biologics in asthma: a molecular perspective to precision medicine. Front Pharmacol 2021;12:793409 [FREE Full text] [doi: 10.3389/fphar.2021.793409] [Medline: 35126131]
36. van der Burg N, Tufvesson E. Is asthma's heterogeneity too vast to use traditional phenotyping for modern biologic therapies? Respir Med 2023;212:107211. [doi: 10.1016/j.rmed.2023.107211] [Medline: 36924848]
37. A study of the qualitative impact of non-medical switching. Alliance for Patient Access. 2019. URL: https://tinyurl.com/2vxwks83 [accessed 2024-01-22]
38. Cost-motivated treatment changes & non-medical switching: commercial health plans analysis. Alliance for Patient Access. 2017. URL: https://tinyurl.com/424dy3xz [accessed 2024-01-22]
39. Collins S. Asthma meds, insurers, and the practice of non-medical drug switching. HealthCentral. 2023. URL: https://www.healthcentral.com/condition/asthma/what-you-need-to-know-about-asthma-meds [accessed 2024-01-22]
40. Landhuis E. OTC budesonide-formoterol for asthma could save lives, money. Medscape Medical News. 2023. URL: https://www.medscape.com/viewarticle/989099 [accessed 2024-01-22]
41. Modglin L. How much do inhalers cost? SingleCare. 2022. URL: https://www.singlecare.com/blog/asthma-inhalers-price-list [accessed 2024-01-22]
42. Gibson PG, McDonald VM, Thomas D. Treatable traits, combination inhaler therapy and the future of asthma management. Respirology 2023;28(9):828-840 [FREE Full text] [doi: 10.1111/resp.14556] [Medline: 37518933]
43. Dahlin A, Denny J, Roden DM, Brilliant MH, Ingram C, Kitchner TE, et al. CMTR1 is associated with increased asthma exacerbations in patients taking inhaled corticosteroids. Immun Inflamm Dis 2015;3(4):350-359 [FREE Full text] [doi: 10.1002/iid3.73] [Medline: 26734457]
44. Keskin O, Farzan N, Birben E, Akel H, Karaaslan C, Maitland-van der Zee AH, et al. Genetic associations of the response to inhaled corticosteroids in asthma: a systematic review. Clin Transl Allergy 2019;9:2 [FREE Full text] [doi: 10.1186/s13601-018-0239-2] [Medline: 30647901]
45. Delgado-Dolset MI, Obeso D, Rodríguez-Coira J, Tarin C, Tan G, Cumplido JA, et al. Understanding uncontrolled severe allergic asthma by integration of omic and clinical data. Allergy 2022;77(6):1772-1785 [FREE Full text] [doi: 10.1111/all.15192] [Medline: 34839541]
46. Liu Q, Hua L, Bao C, Kong L, Hu J, Liu C, et al. Inhibition of spleen tyrosine kinase restores glucocorticoid sensitivity to improve steroid-resistant asthma. Front Pharmacol 2022;13:885053 [FREE Full text] [doi: 10.3389/fphar.2022.885053] [Medline: 35600871]
47. Cardoso-Vigueros C, von Blumenthal T, Rückert B, Rinaldi AO, Tan G, Dreher A, et al. Leukocyte redistribution as immunological biomarker of corticosteroid resistance in severe asthma. Clin Exp Allergy 2022;52(10):1183-1194 [FREE Full text] [doi: 10.1111/cea.14128] [Medline: 35305052]
48. Liang H, Zhang X, Ma Z, Sun Y, Shu C, Zhu Y, et al. Association of CYP3A5 gene polymorphisms and amlodipine-induced peripheral edema in Chinese Han patients with essential hypertension. Pharmgenomics Pers Med 2021;14:189-197 [FREE Full text] [doi: 10.2147/PGPM.S291277] [Medline: 33564260]
49. Wang SB, Huang T. The early detection of asthma based on blood gene expression. Mol Biol Rep 2019;46(1):217-223. [doi: 10.1007/s11033-018-4463-6] [Medline: 30421126]

50. Roberts JK, Moore CD, Romero EG, Ward RM, Yost GS, Reilly CA. Regulation of CYP3A genes by glucocorticoids in human lung cells. F1000Res 2013;2:173 [FREE Full text] [doi: 10.12688/f1000research.2-173.v2] [Medline: 24555085]
51. Moore CD, Roberts JK, Orton CR, Murai T, Fidler TP, Reilly CA, et al. Metabolic pathways of inhaled glucocorticoids by the CYP3A enzymes. Drug Metab Dispos 2013;41(2):379-389 [FREE Full text] [doi: 10.1124/dmd.112.046318] [Medline: 23143891]
52. Roche N, Garcia G, de Larrard A, Cancalon C, Bénard S, Perez V, et al. Real-life impact of uncontrolled severe asthma on mortality and healthcare use in adolescents and adults: findings from the retrospective, observational RESONANCE study in France. BMJ Open 2022;12(8):e060160 [FREE Full text] [doi: 10.1136/bmjopen-2021-060160] [Medline: 36002203]
53. Munoz-Cano R, Torrego A, Bartra J, Sanchez-Lopez J, Palomino R, Picado C, et al. Follow-up of patients with uncontrolled asthma: clinical features of asthma patients according to the level of control achieved (the COAS study). Eur Respir J 2017;49(3):1501885 [FREE Full text] [doi: 10.1183/13993003.01885-2015] [Medline: 28254764]
54. Stockmann C, Reilly CA, Fassl B, Gaedigk R, Nkoy F, Stone B, et al. Effect of CYP3A5*3 on asthma control among children treated with inhaled beclomethasone. J Allergy Clin Immunol 2015;136(2):505-507 [FREE Full text] [doi: 10.1016/j.jaci.2015.02.009] [Medline: 25825214]
55. Stockmann C, Fassl B, Gaedigk R, Nkoy F, Uchida DA, Monson S, et al. Fluticasone propionate pharmacogenetics: CYP3A4*22 polymorphism and pediatric asthma control. J Pediatr 2013;162(6):1222-1227, 1227.e1-2 [FREE Full text] [doi: 10.1016/j.jpeds.2012.11.031] [Medline: 23290512]
56. Smolnikova MV, Kasparov EW, Malinchik MA, Kopylova KV. Genetic markers of children asthma: predisposition to disease course variants. Vavilovskii Zhurnal Genet Selektsii 2023;27(4):393-400 [FREE Full text] [doi: 10.18699/VJGB-23-47] [Medline: 37465198]
57. Kim HK, Kang JO, Lim JE, Ha TW, Jung HU, Lee WJ, et al. Genetic differences according to onset age and lung function in asthma: a cluster analysis. Clin Transl Allergy 2023;13(7):e12282 [FREE Full text] [doi: 10.1002/clt2.12282] [Medline: 37488738]
58. Mohan A, Lugogo NL. Phenotyping, precision medicine, and asthma. Semin Respir Crit Care Med 2022;43(5):739-751. [doi: 10.1055/s-0042-1750130] [Medline: 36220058]
59. Casanova S, Ahmed E, Bourdin A. Definition, phenotyping of severe asthma, including cluster analysis. Adv Exp Med Biol 2023;1426:239-252. [doi: 10.1007/978-3-031-32259-4_11] [Medline: 37464124]
60. Singhal P, Tan ALM, Drivas TG, Johnson KB, Ritchie MD, Beaulieu-Jones BK. Opportunities and challenges for biomarker discovery using electronic health record data. Trends Mol Med 2023;29(9):765-776. [doi: 10.1016/j.molmed.2023.06.006] [Medline: 37474378]
61. Huang SD, Bamba V, Bothwell S, Fechner PY, Furniss A, Ikomi C, et al. Development and validation of a computable phenotype for turner syndrome utilizing electronic health records from a national pediatric network. Am J Med Genet A 2024;194(4):e63495. [doi: 10.1002/ajmg.a.63495] [Medline: 38066696]
62. Blecker S, Schoenthaler A, Martinez TR, Belli HM, Zhao Y, Wong C, et al. Leveraging electronic health record technology and team care to address medication adherence: protocol for a cluster randomized controlled trial. JMIR Res Protoc 2023;12:e47930 [FREE Full text] [doi: 10.2196/47930] [Medline: 37418304]
63. Verhoef PA, Spicer AB, Lopez-Espina C, Bhargava A, Schmalz L, Sims MD, et al. Analysis of protein biomarkers from hospitalized COVID-19 patients reveals severity-specific signatures and two distinct latent profiles with differential responses to corticosteroids. Crit Care Med 2023;51(12):1697-1705. [doi: 10.1097/CCM.0000000000005983] [Medline: 37378460]
64. Hu Y, Huerta J, Cordella N, Mishuris RG, Paschalidis IC. Personalized hypertension treatment recommendations by a data-driven model. BMC Med Inform Decis Mak 2023;23(1):44 [FREE Full text] [doi: 10.1186/s12911-023-02137-z] [Medline: 36859187]
65. Cottrill KA, Rad MG, Ripple MJ, Stephenson ST, Mohammad AF, Tidwell M, et al. Cluster analysis of plasma cytokines identifies two unique endotypes of children with asthma in the pediatric intensive care unit. Sci Rep 2023;13(1):3521 [FREE Full text] [doi: 10.1038/s41598-023-30679-9] [Medline: 36864187]
66. Horne EMF, McLean S, Alsallakh MA, Davies GA, Price DB, Sheikh A, et al. Defining clinical subtypes of adult asthma using electronic health records: analysis of a large UK primary care database with external validation. Int J Med Inform 2023;170:104942 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104942] [Medline: 36529028]
67. Ilmarinen P, Julkunen-Iivari A, Lundberg M, Luukkainen A, Nuutinen M, Karjalainen J, et al. Cluster analysis of Finnish population-based adult-onset asthma patients. J Allergy Clin Immunol Pract 2023;11(10):3086-3096 [FREE Full text] [doi: 10.1016/j.jaip.2023.05.034] [Medline: 37268268]
68. Imoto S, Suzukawa M, Fukutomi Y, Kobayashi N, Taniguchi M, Nagase T, et al. Phenotype characterization and biomarker evaluation in moderate to severe type 2-high asthma. Asian Pac J Allergy Immunol 2023:1-14 [FREE Full text] [doi: 10.12932/AP-021222-1510] [Medline: 37302094]
69. Kim MA, Shin SW, Park JS, Uh ST, Chang HS, Bae DJ, et al. Clinical characteristics of exacerbation-prone adult asthmatics identified by cluster analysis. Allergy Asthma Immunol Res 2017;9(6):483-490 [FREE Full text] [doi: 10.4168/aair.2017.9.6.483] [Medline: 28913987]

70. Matabuena M, Salgado FJ, Nieto-Fontarigo JJ, Álvarez-Puebla MJ, Arismendi E, Barranco P, et al. Identification of asthma phenotypes in the Spanish MEGA cohort study using cluster analysis. Arch Bronconeumol 2023;59(4):223-231 [FREE Full text] [doi: 10.1016/j.arbres.2023.01.007] [Medline: 36732158]
71. Ngo SY, Venter C, Anderson WC3, Picket K, Zhang H, Arshad SH, et al. Clinical features and later prognosis of replicable early-life wheeze clusters from two birth cohorts 12 years apart. Pediatr Allergy Immunol 2023;34(7):e13999 [FREE Full text] [doi: 10.1111/pai.13999] [Medline: 37492911]
72. Zhan W, Wu F, Zhang Y, Lin L, Li W, Luo W, et al. Identification of cough-variant asthma phenotypes based on clinical and pathophysiologic data. J Allergy Clin Immunol 2023;152(3):622-632. [doi: 10.1016/j.jaci.2023.04.017] [Medline: 37178731]
73. Cloutier MM, Akinbami LJ, Salo PM, Schatz M, Simoneau T, Wilkerson JC, et al. Use of national asthma guidelines by allergists and pulmonologists: a national survey. J Allergy Clin Immunol Pract 2020;8(9):3011-3020.e2 [FREE Full text] [doi: 10.1016/j.jaip.2020.04.026] [Medline: 32344187]
74. Vollmer WM, O'Hollaren M, Ettinger KM, Stibolt T, Wilkins J, Buist AS, et al. Specialty differences in the management of asthma. A cross-sectional assessment of allergists' patients and generalists' patients in a large HMO. Arch Intern Med 1997;157(11):1201-1208. [Medline: 9183231]
75. Cloutier MM, Salo PM, Akinbami LJ, Cohn RD, Wilkerson JC, Diette GB, et al. Clinician agreement, self-efficacy, and adherence with the guidelines for the diagnosis and management of asthma. J Allergy Clin Immunol Pract 2018;6(3):886-894.e4 [FREE Full text] [doi: 10.1016/j.jaip.2018.01.018] [Medline: 29408439]
76. Diette GB, Skinner EA, Nguyen TT, Markson L, Clark BD, Wu AW. Comparison of quality of care by specialist and generalist physicians as usual source of asthma care for children. Pediatrics 2001;108(2):432-437. [doi: 10.1542/peds.108.2.432] [Medline: 11483811]
77. Rosman Y, Hornik-Lurie T, Meir-Shafrir K, Lachover-Roth I, Cohen-Engler A, Confino-Cohen R. The effect of asthma specialist intervention on asthma control among adults. World Allergy Organ J 2022;15(11):100712 [FREE Full text] [doi: 10.1016/j.waojou.2022.100712] [Medline: 36440463]
78. Wu AW, Young Y, Skinner EA, Diette GB, Huber M, Peres A, et al. Quality of care and outcomes of adults with asthma treated by specialists and generalists in managed care. Arch Intern Med 2001;161(21):2554-2560 [FREE Full text] [doi: 10.1001/archinte.161.21.2554] [Medline: 11718586]
79. Erickson S, Tolstykh I, Selby JV, Mendoza G, Iribarren C, Eisner MD. The impact of allergy and pulmonary specialist care on emergency asthma utilization in a large managed care organization. Health Serv Res 2005;40(5 Pt 1):1443-1465 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00410.x] [Medline: 16174142]
80. Zeiger RS, Heller S, Mellon MH, Wald J, Falkoff R, Schatz M. Facilitated referral to asthma specialist reduces relapses in asthma emergency room visits. J Allergy Clin Immunol 1991;87(6):1160-1168. [doi: 10.1016/0091-6749(91)92162-t] [Medline: 2045618]
81. Mahr TA, Evans R3. Allergist influence on asthma care. Ann Allergy 1993;71(2):115-120. [Medline: 8346862]
82. Schatz M, Zeiger RS, Mosen D, Apter AJ, Vollmer WM, Stibolt TB, et al. Improved asthma outcomes from allergy specialist care: a population-based cross-sectional analysis. J Allergy Clin Immunol 2005;116(6):1307-1313 [FREE Full text] [doi: 10.1016/j.jaci.2005.09.027] [Medline: 16337464]
83. Wechsler ME. Managing asthma in primary care: putting new guideline recommendations into context. Mayo Clin Proc 2009;84(8):707-717 [FREE Full text] [doi: 10.4065/84.8.707] [Medline: 19648388]
84. Cooper S, Rahme E, Tse SM, Grad R, Dorais M, Li P. Are primary care and continuity of care associated with asthma-related acute outcomes amongst children? A retrospective population-based study. BMC Prim Care 2022;23(1):5 [FREE Full text] [doi: 10.1186/s12875-021-01605-7] [Medline: 35172739]
85. Akinbami LJ, Salo PM, Cloutier MM, Wilkerson JC, Elward KS, Mazurek JM, et al. Primary care clinician adherence with asthma guidelines: the National Asthma Survey of Physicians. J Asthma 2020;57(5):543-555 [FREE Full text] [doi: 10.1080/02770903.2019.1579831] [Medline: 30821526]
86. HEDIS measures and technical resources: asthma medication ratio (AMR). NCQA. 2023. URL: https://www.ncqa.org/hedis/measures/medication-management-for-people-with-asthma-and-asthma-medication-ratio [accessed 2024-01-22]
87. Schatz M, Zeiger RS, Vollmer WM, Mosen D, Mendoza G, Apter AJ, et al. The controller-to-total asthma medication ratio is associated with patient-centered as well as utilization outcomes. Chest 2006;130(1):43-50. [doi: 10.1378/chest.130.1.43] [Medline: 16840381]
88. Kim Y, Parrish KM, Pirritano M, Moonie S. A higher asthma medication ratio (AMR) predicts a decrease in ED visits among African American and Hispanic children. J Asthma 2023;60(7):1428-1437. [doi: 10.1080/02770903.2022.2155183] [Medline: 36461904]
89. Luskin AT, Antonova EN, Broder MS, Chang E, Raimundo K, Solari PG. Patient outcomes, health care resource use, and costs associated with high versus low HEDIS asthma medication ratio. J Manag Care Spec Pharm 2017;23(11):1117-1124 [FREE Full text] [doi: 10.18553/jmcp.2017.23.11.1117] [Medline: 29083971]
90. Andrews AL, Simpson AN, Basco WTJ, Teufel RJ2. Asthma medication ratio predicts emergency department visits and hospitalizations in children with asthma. Medicare Medicaid Res Rev 2013;3(4):mmrr.003.04.a05 [FREE Full text] [doi: 10.5600/mmrr.003.04.a05] [Medline: 24834366]

91. Andrews AL, Brinton DL, Simpson KN, Simpson AN. A longitudinal examination of the asthma medication ratio in children with Medicaid. J Asthma 2020;57(10):1083-1091 [FREE Full text] [doi: 10.1080/02770903.2019.1640727] [Medline: 31313611]
92. Andrews AL, Brinton D, Simpson KN, Simpson AN. A longitudinal examination of the asthma medication ratio in children. Am J Manag Care 2018;24(6):294-300 [FREE Full text] [Medline: 29939504]
93. Tong Y, Messinger AI, Wilcox AB, Mooney SD, Davidson GH, Suri P, et al. Forecasting future asthma hospital encounters of patients with asthma in an academic health care system: predictive model development and secondary analysis study. J Med Internet Res 2021;23(4):e22796 [FREE Full text] [doi: 10.2196/22796] [Medline: 33861206]
94. Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. JMIR Med Inform 2020;8(1):e16080 [FREE Full text] [doi: 10.2196/16080] [Medline: 31961332]
95. Luo G, Nau CL, Crawford WW, Schatz M, Zeiger RS, Rozema E, et al. Developing a predictive model for asthma-related hospital encounters in patients with asthma in a large, integrated health care system: secondary analysis. JMIR Med Inform 2020;8(11):e22689 [FREE Full text] [doi: 10.2196/22689] [Medline: 33164906]
96. Mosen DM, Macy E, Schatz M, Mendoza G, Stibolt TB, McGaw J, et al. How well do the HEDIS asthma inclusion criteria identify persistent asthma? Am J Manag Care 2005;11(10):650-654 [FREE Full text] [Medline: 16232006]
97. Schatz M, Zeiger RS. Improving asthma outcomes in large populations. J Allergy Clin Immunol 2011;128(2):273-277 [FREE Full text] [doi: 10.1016/j.jaci.2011.03.027] [Medline: 21497885]
98. Schatz M, Zeiger RS, Yang SJ, Chen W, Crawford WW, Sajjan SG, et al. Persistent asthma defined using HEDIS versus survey criteria. Am J Manag Care 2010;16(11):e281-e288 [FREE Full text] [Medline: 21087074]
99. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004;10(1):25-32 [FREE Full text] [Medline: 14738184]
100. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003;9(8):538-547 [FREE Full text] [Medline: 12921231]
101. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998;157(4 Pt 1):1173-1180 [FREE Full text] [doi: 10.1164/ajrccm.157.4.9708124] [Medline: 9563736]
102. Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty of half full? J Asthma 1999;36(4):359-370. [doi: 10.3109/02770909909068229] [Medline: 10386500]
103. Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, et al. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010;138(5):1156-1165 [FREE Full text] [doi: 10.1378/chest.09-2426] [Medline: 20472862]
104. Loymans RJB, Debray TPA, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TRJ, et al. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-1952.e15 [FREE Full text] [doi: 10.1016/j.jaip.2018.02.004] [Medline: 29454163]
105. Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest 2012;141(1):58-65. [doi: 10.1378/chest.11-0020] [Medline: 21885725]
106. Sato R, Tomita K, Sano H, Ichihashi H, Yamagata S, Sano A, et al. The strategy for predicting future exacerbation of asthma using a combination of the asthma control test and lung function test. J Asthma 2009;46(7):677-682. [doi: 10.1080/02770900902972160] [Medline: 19728204]
107. Yurk RA, Diette GB, Skinner EA, Dominici F, Clark RD, Steinwachs DM, et al. Predicting patient-reported asthma outcomes for adults in managed care. Am J Manag Care 2004;10(5):321-328 [FREE Full text] [Medline: 15152702]
108. Xiang Y, Ji H, Zhou Y, Li F, Du J, Rasmy L, et al. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: retrospective cohort study. J Med Internet Res 2020;22(7):e16981 [FREE Full text] [doi: 10.2196/16981] [Medline: 32735224]
109. Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, et al. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006;28(6):1145-1155 [FREE Full text] [doi: 10.1183/09031936.06.00145105] [Medline: 16870656]
110. Loymans RJ, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJ, Schermer TR, et al. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016;71(9):838-846 [FREE Full text] [doi: 10.1136/thoraxjnl-2015-208138] [Medline: 27044486]
111. Schatz M. Predictors of asthma control: what can we modify? Curr Opin Allergy Clin Immunol 2012;12(3):263-268. [doi: 10.1097/ACI.0b013e32835335ac] [Medline: 22517290]
112. Dick S, Doust E, Cowie H, Ayres JG, Turner S. Associations between environmental exposures and asthma control and exacerbations in young children: a systematic review. BMJ Open 2014;4(2):e003827 [FREE Full text] [doi: 10.1136/bmjopen-2013-003827] [Medline: 24523420]

113. Schwartz J, Slater D, Larson TV, Pierson WE, Koenig JQ. Particulate air pollution and hospital emergency room visits for asthma in Seattle. Am Rev Respir Dis 1993;147(4):826-831. [doi: 10.1164/ajrccm/147.4.826] [Medline: 8466116]
114. Romieu I, Meneses F, Sienra-Monge JJ, Huerta J, Ruiz Velasco S, White MC, et al. Effects of urban air pollutants on emergency visits for childhood asthma in Mexico City. Am J Epidemiol 1995;141(6):546-553. [doi: 10.1093/oxfordjournals.aje.a117470] [Medline: 7900722]
115. Lu P, Zhang Y, Lin J, Xia G, Zhang W, Knibbs LD, et al. Multi-city study on air pollution and hospital outpatient visits for asthma in China. Environ Pollut 2020;257:113638. [doi: 10.1016/j.envpol.2019.113638] [Medline: 31812526]
116. Liu Y, Pan J, Zhang H, Shi C, Li G, Peng Z, et al. Short-term exposure to ambient air pollution and asthma mortality. Am J Respir Crit Care Med 2019;200(1):24-32 [FREE Full text] [doi: 10.1164/rccm.201810-1823OC] [Medline: 30871339]
117. Vagaggini B, Taccola M, Cianchetti S, Carnevali S, Bartoli ML, Bacci E, et al. Ozone exposure increases eosinophilic airway response induced by previous allergen challenge. Am J Respir Crit Care Med 2002;166(8):1073-1077 [FREE Full text] [doi: 10.1164/rccm.2201013] [Medline: 12379550]
118. Sanchez P, Voisey JP, Xia T, Watson HI, O'Neil AQ, Tsaftaris SA. Causal machine learning for healthcare and precision medicine. R Soc Open Sci 2022;9(8):220638 [FREE Full text] [doi: 10.1098/rsos.220638] [Medline: 35950198]
119. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model 1986;7(9-12):1393-1512 [FREE Full text] [doi: 10.1016/0270-0255(86)90088-6]
120. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol 2011;173(7):731-738 [FREE Full text] [doi: 10.1093/aje/kwq472] [Medline: 21415029]
121. Efron B. Bootstrap methods: another look at the Jackknife. Ann Stat 1979;7(1):1-26 [FREE Full text] [doi: 10.1214/aos/1176344552]
122. Nkoy FL, Stone BL, Fassl BA, Uchida DA, Koopmeiners K, Halbern S, et al. Longitudinal validation of a tool for asthma self-monitoring. Pediatrics 2013;132(6):e1554-e1561 [FREE Full text] [doi: 10.1542/peds.2013-1389] [Medline: 24218469]
123. Anghel LA, Farcas AM, Oprean RN. An overview of the common methods used to measure treatment adherence. Med Pharm Rep 2019;92(2):117-122 [FREE Full text] [doi: 10.15386/mpr-1201] [Medline: 31086837]

Abbreviations

AMR: asthma medication ratio
CMTR1: cap methyltransferase 1
CYP: cytochrome P
EHR: electronic health record
ER: emergency room
FeNO: fractional exhaled nitric oxide
HEDIS: Healthcare Effectiveness Data and Information Set
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
***ICD-10*:** *International Classification of Diseases, Tenth Revision*
ICS: inhaled corticosteroid
IgE: immunoglobulin E
IL: interleukin
MAGI2: membrane associated guanylate kinase, WW and PDZ domain containing 2
mHealth: mobile health
PCP: primary care provider
Th2: T-helper type 2
TRIM24: tripartite motif containing 24

Edited by A Benis; submitted 24.01.24; peer-reviewed by H Tibble, A Kaplan; comments to author 01.03.24; revised version received 12.03.24; accepted 25.03.24; published 17.04.24.

Please cite as:
Nkoy FL, Stone BL, Zhang Y, Luo G
A Roadmap for Using Causal Inference and Machine Learning to Personalize Asthma Medication Selection
JMIR Med Inform 2024;12:e56572
URL: https://medinform.jmir.org/2024/1/e56572
doi:10.2196/56572
PMID:38630536

Viewpoint

AI: Bridging Ancient Wisdom and Modern Innovation in Traditional Chinese Medicine

Linken Lu[1]; Tangsheng Lu[2]; Chunyu Tian[1]; Xiujun Zhang[3], PhD

[1]North China University of Science and Technology, Tangshan, China

[2]National Institute on Drug Dependence and Beijing Key Laboratory of Drug Dependence Research, Peking University, Beijing, China

[3]School of Psychology and Mental Health, North China University of Science and Technology, Hebei Province, China

Corresponding Author:
Xiujun Zhang, PhD
School of Psychology and Mental Health
North China University of Science and Technology
21 Bohai Avenue, Caofeidian New Town, Tangshan
Hebei Province, 063210
China
Phone: 86 0315 8805970
Email: zhxj@ncst.edu.cn

Abstract

The pursuit of groundbreaking health care innovations has led to the convergence of artificial intelligence (AI) and traditional Chinese medicine (TCM), thus marking a new frontier that demonstrates the promise of combining the advantages of ancient healing practices with cutting-edge advancements in modern technology. TCM, which is a holistic medical system with >2000 years of empirical support, uses unique diagnostic methods such as inspection, auscultation and olfaction, inquiry, and palpation. AI is the simulation of human intelligence processes by machines, especially via computer systems. TCM is experience oriented, holistic, and subjective, and its combination with AI has beneficial effects, which presumably arises from the perspectives of diagnostic accuracy, treatment efficacy, and prognostic veracity. The role of AI in TCM is highlighted by its use in diagnostics, with machine learning enhancing the precision of treatment through complex pattern recognition. This is exemplified by the greater accuracy of TCM syndrome differentiation via tongue images that are analyzed by AI. However, integrating AI into TCM also presents multifaceted challenges, such as data quality and ethical issues; thus, a unified strategy, such as the use of standardized data sets, is required to improve AI understanding and application of TCM principles. The evolution of TCM through the integration of AI is a key factor for elucidating new horizons in health care. As research continues to evolve, it is imperative that technologists and TCM practitioners collaborate to drive innovative solutions that push the boundaries of medical science and honor the profound legacy of TCM. We can chart a future course wherein AI-augmented TCM practices contribute to more systematic, effective, and accessible health care systems for all individuals.

(JMIR Med Inform 2024;12:e58491) doi:10.2196/58491

KEYWORDS

traditional Chinese medicine; TCM; artificial intelligence; AI; diagnosis

Introduction

Traditional Chinese medicine (TCM) is a vibrant and enduring medical system that has been refined for thousands of years, thus offering a rich tapestry of health and healing practices [1]. With its roots deeply embedded in Chinese philosophy and a profound understanding of the human body's relationship with the natural world, TCM has developed a unique set of diagnostic and therapeutic methodologies. These methodologies include herbal medicine, acupuncture, and the identification of syndrome patterns, which are grounded in the fundamental concepts of yin and yang and the 5-element theory, thus providing a holistic approach to health that addresses the body, mind, and spirit. As a traditional medical system, TCM primarily depends on personal experience and lacks standardized and systematic diagnosis and treatment procedures, which potentially discourages its more widespread adoption. Thus, the rapid expansion of artificial intelligence (AI) can significantly improve the reliability and accuracy of TCM diagnostics, thereby increasing the effective use of therapeutic methods for patients [2]. AI is currently being used in diagnostics to detect abnormalities in medical imaging, such as identifying lung nodules or other suspicious lesions in early cancer screening

[3]. These tools can assist physicians in making an initial diagnosis by analyzing large amounts of imaging data to identify potential signs of cancer and can also help physicians more accurately characterize suspicious lesions, including their shape, volume, histopathological diagnosis, disease stage, and molecular features. AI also has potential therapeutic value for the diagnosis of mental disorders such as major depressive disorder by successfully distinguishing patients from healthy controls through machine learning [4,5]. This characterization is critical for determining treatment options and predicting disease progression. However, applying AI to TCM exhibits a range of both opportunities and challenges. AI can enhance the diagnostic process by enabling physicians to make more accurate assessments of diseases and patient constitutions through the analysis of extensive TCM clinical data. It can also be instrumental in creating personalized treatment plans that are tailored to the unique conditions of each patient. Concurrently, the integration of AI into TCM raises significant concerns about patient privacy and data security. Although the establishment and learning of classification criteria still cannot eliminate subjectivity, efforts must be made to refine these processes. Addressing these issues necessitates the establishment of stringent ethical standards and robust privacy measures.

The successful integration of AI into TCM is contingent upon collaborative efforts across various disciplines, including medicine, computer science, and data science. Building interdisciplinary teams and fostering effective communication are crucial for driving innovation in this field. As technology evolves, it possesses the potential to revolutionize TCM practices provided that it is implemented with careful consideration of the ethical implications and needs of the TCM community. In the realm of TCM, diagnosis is a pivotal element wherein practitioners engage in comprehensive inquiry and conduct head-to-toe examinations of patients. This process is central to collecting health-related data. Such a diagnostic approach facilitates an in-depth evaluation of the patient's overall health status coupled with a nuanced understanding of the fundamental characteristics of the disease. TCM diagnosis is based on a holistic perspective and involves a comprehensive assessment of both physiological and psychological aspects of health rather than relying solely on objective diagnostic criteria such as molecular markers and physiological indicators. There are generally 4 main diagnostic methods in TCM: inspection, auscultation and olfaction, inquiry, and palpation. These facets of TCM have been widely accepted by TCM practitioners worldwide. This aspect of TCM is particularly pronounced for highly skilled physicians who are able to derive diagnostic insights through visual inspection alone, often without an explicit need for detailed procedural explication. Moreover, they believe that nonrational thinking, which encompasses implicit meanings, intuition, inspiration, and imagination, plays a vital role in TCM diagnosis and treatment [6]. In conjunction with the concept of the 4 diagnostic methods, practitioners assess and integrate various clinical data to investigate evidence, identify a disease's root causes, establish treatment strategies, assess treatment efficacy, and anticipate healing progress [7]. The integration of AI into TCM diagnosis respects and uses the intuition and experience of practitioners, thus serving as an auxiliary means of clinical research to evaluate and verify diagnostic results rather than replace human judgment. AI can be designed to analyze complex patterns in patient data, thereby augmenting traditional diagnostic methods. For example, machine learning models can be trained to recognize subtleties in patient inquiry responses, thus enhancing the practitioner's ability to identify the root causes of diseases and tailor treatment strategies. AI systems can be developed to consider the holistic approach that is inherent in TCM. By analyzing a wide range of clinical data, AI can provide a more comprehensive health assessment that aligns with the TCM concept of integrating various aspects of a patient's condition. This not only supports the practitioner's diagnostic process but also aids in anticipating healing progress and the efficacy of treatments. In addition, AI can tailor treatments to patients' individual needs and conditions by considering their unique body states and responses to various therapeutic interventions, thus leveraging the ability of AI to process and learn from large data sets as well as the enormous potential for personalized treatment in TCM. The role of AI in TCM is to augment the expertise of physicians, thus providing insights and analytics that support the holistic and personalized approach to health care that is at the heart of TCM. By doing so, AI can contribute to enhancing patient care and ensuring that the rich heritage of TCM is carried forward and developed in the modern health care era.

AI in TCM Diagnosis

Inspection

In the practice of TCM, the method of inspection primarily focuses on the acquisition of diagnostic information through direct observation. This approach involves assessing the patient's condition by scrutinizing various physical changes across the body. Distinct from the paradigm of Western medicine, which predominantly relies on objective, empirical evidence, TCM tends to base its diagnostic conclusions on subjective interpretations by medical practitioners [8]. The scope of inspection for diagnosis is quite broad. Although the methods comprise craniofacial observations, tongue and face diagnoses are the primary methods used for inspection. An inspection of the tongue's shape, size, color, and texture aids in the assessment of organ function and the development of medical conditions. Facial expression analysis is a diagnostic method that aligns with the theory of 5 zang organs, corresponding to 5 elements and colors. It involves distinguishing different changes in facial color, such as green, red, yellow, white, and black, based on the principles of yin and yang and the 5-element theory [9]. Building on this traditional foundation, technological advancements have introduced new methods to enhance TCM diagnostics. For example, the development by Chen [10] of a neural network–based system marked a significant advancement in this research. This innovative platform automates the diagnostic and treatment process in TCM with a focus on symptom analysis. It empowers physicians to efficiently access crucial medical records, thus providing valuable insights into the effectiveness of TCM treatments for similar conditions. Furthermore, the system streamlines the prescription process, thus enabling precise electronic prescriptions to be quickly dispatched to relevant departments and patients. This integration of AI with TCM improves diagnostic accuracy and facilitates

the sharing of expertise among practitioners, thus ultimately enhancing the standard of care in TCM. AI is currently one of the most discussed topics in medical imaging research. It serves as a significant enabler for handling vast amounts of medical images, thereby deciphering disease features that may be imperceptible to the human eye. Similarly, AI-based facial diagnosis and tongue diagnosis hold promise for further development. Recently, Liu et al [11] reviewed AI methods in the field of tongue diagnosis. They identified two main challenges that hinder development in this field: (1) the authority of data sets and (2) a misconception about a sole reliance on single features for diagnosis in traditional Chinese tongue diagnosis [11]. The combination of AI with this field overcomes the inherent subjectivity of TCM diagnosis and provides a more objective and standardized approach to tongue diagnosis. Technological advances such as multiscale features and the incorporation of previous knowledge have been successfully applied to improve the accuracy and reliability of AI-assisted tongue analysis. In addition, robust data sets and reliable performance evaluations are still needed to address existing problems in the field. The future of intelligent tongue diagnosis is promising, with potential breakthroughs in self-supervised methods, multimodal information fusion, and tongue pathology research that are expected to have a significant impact on research and clinical practice. On the basis of this scenario, we propose potential solutions to address these issues. First, standardizing data sets for tongue diagnosis should be a collaborative effort that involves experts in TCM. Second, leveraging multimodal data in AI is a crucial approach for the AI-driven transformation of TCM.

The future of AI in the inspection component of TCM is poised to transform traditional diagnostic practices through innovative research and practical applications. One of the primary research directions is the development of sophisticated AI algorithms that can analyze and interpret tongue and facial diagnostics at a level of detail that surpasses human perception. By training these algorithms on diverse and high-quality data sets, AI systems can learn to identify subtle patterns and changes that indicate underlying health conditions, thus complementing the expertise of TCM practitioners.

Auscultation and Olfaction

Auscultation and olfaction in TCM involve the use of a physician's hearing to detect changes in a patient's voice and sounds. Olfaction relies on the physician's sense of smell to detect changes in odors. The theoretical basis for these practices in TCM is the belief that a patient's speech sounds and body odors can reflect the physiological and psychological states of their internal organs. Consequently, auscultation and olfaction have long been highly regarded in the field of TCM. However, objective studies and literature on auscultation and olfaction are scarce, which may be attributed to the complex acoustic properties of sounds, including a plethora of natural noises, similar acoustic signals, and diverse chemical compositions of thousands of volatile organic compounds in exhaled gases. These factors have hindered the development of objective research on TCM auscultation and olfaction. Chiu et al [12] introduced quantifiable parameters for TCM auscultation, which allowed for the identification of nonvacuity, qi vacuity, and yin vacuity characteristics in participants. This quantification process enhances the practice of TCM auscultation. There is still a need for more quantitative data on auscultation and further advancements in the application of AI to analyze such quantitative data. The integration of AI into this method is facilitated through the use of advanced sensor technologies and audio analysis algorithms. For example, digital stethoscopes can capture and record bodily sounds with greater clarity. These sounds are then processed by AI algorithms that can filter out background noise and enhance the relevant audio signals. With regard to objective olfactory analysis, there have been several recent studies from a TCM perspective. A recent study introduced a new odor map with the ability to characterize odor quality that was comparable to that of highly skilled human "sniffers" [13]. The algorithms of these odor detection tools have significant potential for quantifying olfactory diagnosis in TCM. AI algorithms are then applied to data generated by these devices to identify specific volatile organic compound profiles that are associated with different health conditions. This is a complex task given the vast number of potential volatile organic compounds and their concentrations; however, machine learning models have shown the ability to handle this complexity and provide objective data for diagnosis. Therefore, the primary focus should be on building an AI odor monitoring system. Such a detection system can be developed by selecting specific biomarker reagents [14]. The development of diagnostic molecular biomarkers for olfaction diagnosis in TCM is also a substantial task. These biomarkers should be capable of quantifying the olfactory diagnostic process in TCM more accurately.

However, there are still some challenges in the application of AI to auscultation and olfaction, including challenges regarding data quality and standardization. The collection of auditory and olfactory data requires highly accurate sensors and devices. Data quality and standardization are critical for training accurate AI models. Inaccurate or inconsistent data can lead to misjudgments by the AI system. The second challenge involves the recognition of extremely complex sound and smell patterns that can be perceived differently among individuals. AI needs to be able to recognize and understand these complexities, which places high demands on the design and training of algorithms. We need to explore and develop more advanced sensor technologies to improve the accuracy and consistency of data collection and use deep learning techniques to improve the ability of AI models to recognize complex sound and odor patterns.

Inquiry

Interrogation diagnosis (or inquiry diagnosis) directly asks patients questions about various physiological and psychological feelings. This methodology includes gathering information about the patient's family history, primary complaints, living conditions, dietary habits, sleep patterns, and other physical condition characteristics. This process allows the practitioner to gain a comprehensive understanding of the patient's overall health, including factors that may contribute to their current condition. A thorough understanding of a patient can also avoid the influence of a previous medical history on treatment. The inquiry aims to provide a holistic view of the patient in

consideration of not only physical symptoms but also lifestyle and environmental factors that could impact their health. The content of TCM inquiries is mainly based on the "Ten Brief Inquiries"; however, at present, TCM inquiries also incorporate past history, allergy history, and family history from modern medical records [15]. The GatorTron system, which was developed by Yang et al [16], enhances the use of clinical narratives in the creation of various medical AI systems, thus ultimately leading to better health care delivery and health outcomes. However, electronic health record (EHR) analysis for TCM inquiry is not yet well developed and primarily relies on natural language extraction techniques to extract electronic medical record data, which are then used to establish a knowledge repository for traditional Chinese clinical cases. For example, AI systems are capable of identifying TCM-specific symptoms such as "fatigue" and "dry mouth" from patient narratives, thus correlating these symptoms with associated internal organ imbalances. This sophisticated recognition aids physicians in assessing patients' constitutions and developing personalized treatment plans. For individuals with chronic conditions, AI facilitates a more in-depth analysis by sifting through extensive health records to forecast disease progression, thereby providing physicians with a solid foundation for accurate diagnoses. Moreover, AI extends its support to patients who require ongoing care by offering tailored advice on diet and exercise, thus significantly contributing to the enhancement of their quality of life and the mitigation of relapse risks. The advent of smart wearables has further empowered AI by enabling real-time health data collection, which is swiftly relayed to AI for analysis. This system proactively notifies health care providers and patients about emerging health concerns, thus exemplifying the potential of AI in diagnostics and proactive patient care within the framework of TCM.

In the future, it will be essential to confirm the accuracy of large language models (LLMs) such as GPT-3.5 and GPT-4 in TCM diagnosis [17]. This process requires a nuanced approach that acknowledges the complexity and richness of TCM terminology. The first step is to collect comprehensive patient data, including symptoms, medical history, lifestyle factors, and any other relevant information. These data must be preprocessed to ensure that they are suitable for AI analysis. AI models, especially LLMs, are trained in neurolinguistic programming to understand and interpret human language. In the context of TCM, this involves training models to recognize and analyze specific terminology that is used in patient inquiries. Afterward, AI models must be trained to understand the context in which TCM terms are used. This includes recognizing relationships between different symptoms and their implications with regard to overall health according to TCM principles. However, TCM is practiced worldwide, and patient inquiries may also be in various languages or dialects. AI models need to be trained on diverse data sets to ensure that they can handle different languages and cultural interpretations of TCM terms. A significant amount of labeled data and expert input are subsequently required for validation. Collaborations with TCM practitioners to annotate and validate data can improve the model's accuracy. In summary, although AI with LLMs shows significant promise for managing EHRs, TCM inquiry demonstrates a unique knowledge system. The fine-tuning of LLMs is essential for transforming these general-purpose models into specialized models that are adept at handling TCM EHRs [18]. Future efforts should entail constructing a knowledge system for TCM diagnosis. It will then be necessary to fine-tune LLMs for TCM diagnosis based on existing LLM data models, thus providing AI tools for case analysis in TCM diagnosis. The integration of AI into the TCM inquiry process is a complex task that requires the careful consideration of unique aspects of TCM terminology and practice. With the right approach, including ongoing research and collaboration with TCM experts, AI can be effectively used to analyze patient data and enhance the diagnostic process in TCM.

Palpation

Pulse diagnosis is one of the 4 main pillars of TCM assessment. By palpating the pulse at 3 specific positions on the wrists ("cun," "guan," and "chi"), practitioners can gain a comprehensive understanding of a person's overall health and the state of specific organs. TCM pulse diagnosis consists of approximately 29 different pulse types that encompass a range of descriptors, including floating pulses and scattered pulses [19]. The intersection of pulse diagnosis and AI presents 2 main challenges. TCM pulse detection has historically relied on manually palpating the arteries beneath the skin to detect the pulse, thus lacking objective standards. In the process of AI-driven traditional Chinese pulse diagnosis, 2 critical issues need to be addressed: the development of pulse measurement devices and the standardization of pulse detection data. Lan et al [20] created a sensing device that features a multipoint sensor to measure pulse. Due to the complexity of pulse detection, previous methods that have primarily relied on multipoint sensors have only offered a limited scope of information. The development of pulse measurement devices has led to significant technological advances in recent years, and these advances are mainly reflected in innovations in sensor technology and the application of AI algorithms. Photovoltaic volumetric pulse wave sensors, which are based on photoplethysmography, are among the most common types of sensors used in pulse measurement devices. Pulse waves are measured by detecting the flow of blood in the microvasculature to obtain physiological parameters such as heart rate [21]. Photoplethysmographic sensors, such as smartwatches and fitness trackers, are widely used in consumer electronics. Some devices use pressure sensors to measure pulse waves, especially in continuous blood pressure monitoring. These sensors are often embedded in wearable devices that can monitor changes in blood pressure over time. To address the challenge of normalizing pulse data, AI algorithms preprocess the data before analysis, including filtering, denoising, and normalization, to ensure data quality. AI technology that is currently under development is working to improve the cross-device compatibility of algorithms so that data from devices from different manufacturers and models can be consistently analyzed, thus promoting data standardization and interoperability. The standardization of TCM pulse diagnosis is key to promoting the use of AI technology in TCM pulse measurements. It is necessary to establish unified pulse data and diagnostic standards along with integrating more diagnostic methods such as tongue diagnosis and diagnostic observation, from which we can develop an integrated TCM

diagnostic platform and improve the comprehensiveness and accuracy of diagnosis.

In the future, more powerful multipoint sensing devices and multimodal detection devices will be needed to comprehensively examine pulse data and achieve better quantification. A challenge still remains in determining whether pulse data from these detectors can adequately reflect the characteristics of pulse diagnosis and in improving the classification of pulse patterns. To address the challenge of enhancing the precision of AI in interpreting pulse data for future research and development, noncontact pulse measurement techniques have demonstrated significant advancements. These methods eliminate the need for physical contact with the patient, which is particularly crucial for monitoring in unique or sensitive situations. For instance, the polarized multispectral imaging technique for noncontact heart rate measurement has refined the accuracy of data acquisition by pioneering new methods for extracting pulse waves from the palm [22]. This innovation contributes to the establishment of a standardized framework for pulse data, thus facilitating seamless data sharing and comparisons across various devices and systems. However, the attainment of high-quality data hinges on precise labeling, which is a process that can be both labor intensive and costly. In the context of electrocardiogram data annotation, the requirement for specialized physicians introduces variability, in which different medical professionals may offer conflicting assessments. This reality compounds the complexity and challenges associated with data preprocessing. To overcome these obstacles, it is imperative to refine data annotation protocols and invest in the development of more efficient and accurate labeling tools. By doing so, we can ensure that AI systems are trained on the most reliable data, thereby improving their diagnostic capabilities and contributing to the advancement of AI in health care. In summary, the development of detection methods and quantification of detection-based data are bottlenecks in the process of AI-driven pulse diagnosis.

AI-Powered Tuina Massage Robot

Tuina massage (also known as Chinese medical massage) is a traditional hands-on manipulation treatment that is guided by the principles of TCM. It is widely used to treat various ailments, such as knee osteoarthritis, chronic neck pain, and insomnia [23-25]. The tuina massage serves 3 primary functions: facilitating the circulation of meridians, harmonizing qi and blood circulation, and augmenting the immune system [26-28]. These functions are essential for disease prevention and treatment and overall well-being. The integration of AI into tuina massage therapy is in its early stages. Efforts are underway to develop highly intelligent massage equipment and robotics based on TCM tuina to enhance its effectiveness, with a focus on improving the comfort, intelligence, and safety of massage robots [29]. Vibration and percussion are the 2 main types of tuina massage robotics. These devices offer acupressure techniques; however, manual massage from experienced physiotherapists provides additional popular movements, such as light stroking, stretching, and advanced kneading techniques, that machines cannot replicate. Therefore, the development of massage robots is a significant research focus for greater health care demands. Challenges mainly exist in their control, structure, and path planning; however, ongoing efforts aim to optimize their design and functionality. For example, the incorporation of a series-parallel hybrid structure may enhance flexibility while maintaining stiffness and precision [30]. Future research should focus on ergonomics to design high-performance massage robots that integrate advanced AI technologies for better control, sensing, and essential functions.

In current TCM tuina practice, the Expert Manipulative Massage Automation (EMMA) electronic massager, which was developed by AiTreat Pte Ltd in Singapore, is widely used. To deliver precise and effective massage based on muscle feedback, EMMA uses advanced sensor-based technology to identify focus points and adjust pressure levels. By detecting stiffness and resistance, EMMA can pinpoint muscle knots and tension points, thus applying varying pressure levels based on feedback and user preferences. In addition, EMMA incorporates Internet of Things technology for remote control, programming, and updates, thus enhancing its functionality in “green” Internet of Things applications. In EMMA technology, machine learning algorithms (especially convolutional neural networks in deep learning) are used to identify and analyze muscle tension patterns, acupuncture point locations, and physiological responses of patients. Through training, these algorithms are able to identify specific treatment points from sensor data to provide a customized massage solution for the patient. This pattern recognition capability allows the robotic massage therapist to pinpoint the area to be treated, thus mimicking the diagnostic process of an experienced massage therapist. The robotic masseur is able to adjust the intensity and speed of the massage based on real-time feedback from the patient. For example, if the sensors detect that a patient is experiencing discomfort at a certain pressure level, then the AI system can immediately adjust the intensity to ensure the comfort and effectiveness of the treatment. Through its advanced data analytics and learning capabilities, the AI application in EMMA technology is able to accurately identify treatment points and adjust massage intensity based on the patient’s real-time feedback. The EMMA massager has garnered high levels of acceptability and satisfaction among healthy volunteers, thus demonstrating its feasibility [31]. Nonetheless, research on massage robots still faces challenges, particularly regarding their clinical effectiveness. In addition, massage robots are categorized as class-I medical devices that do not require Food and Drug Administration approval for marketing in the United States [32]. Traditional medical device classification focuses on physical and biological characteristics, whereas the functionality of AI devices relies more on software and algorithms. Therefore, new classification criteria need to be developed that consider the specificities and potential risks of AI technologies. In the future, there will be a need for more standardized regulations to oversee research on massage robots [33]. Medical devices process and analyze large amounts of patient data, which requires regulations to include stringent requirements for data security and privacy protection. Medical device regulations need to incorporate specifications for data collection, storage, processing, and transmission to ensure the security and confidentiality of patients’ information, including the validation and clinical testing of AI algorithms. The use of AI medical devices involves the collection and analysis of large

amounts of personal health information, which can threaten patients' privacy. Regulations must ensure that the collection and use of patient information comply with privacy protection standards to prevent unauthorized access and data breaches. This will entail validating the functionality and therapeutic efficacy of medical devices to guarantee their safety and efficacy for users.

We propose the following perspective on the development of tuina robots. First, the overall stability of the tuina technique involves the stability of variable mechanical parameters and resulting morphological changes in mechanical effects during technique operation. These factors include mechanical characteristics such as force; speed; frequency; displacement; and kinematic features such as limb range of motion, joint angles, and overall movement amplitude. For example, the dexterity of the robotic arm is key to achieving an accurate simulation of a human masseur's maneuvers; however, it requires sophisticated mechanical design, including joint flexibility, end-effector versatility, and overall structural stability. The robot arm's control system also needs to process large amounts of data and make decisions in real time. This includes trajectory planning, motion control, and complex algorithms for force and position control. To ensure safety, collision detection and response mechanisms also need to be implemented. Consequently, tuina has significant limitations and subjectivity, thus making it difficult to objectively quantify and accurately assess efficacy. AI offers unique advantages in addressing this issue, which is primarily manifested in the digitization of tuina techniques (ie, the development of precision and flexibility in massage robots). Addressing the accuracy of the tuina technique is a prominent issue that may require more diverse AI algorithms to digitize massage techniques and analyze the clinical effects of different massage methods. The accuracy of Chinese massage largely depends on the precise positioning of acupoints. Researchers are developing a human body model based on the mechanism of "bone degree and minutes" in Chinese medicine, which realizes the calculation of 3D coordinate values of acupoints through robotics as well as identifying and tracking human body features by using such sensor technologies as depth cameras, thus realizing the precise positioning of acupoints. Second, another advantage of AI includes personalized health care services. The personalized parameter settings of massage robots are core parameters for future tuina robots. There are significant differences in individuals' sensitivity and tolerance to pressure. The comfort and pain thresholds of people can vary, thus significantly affecting their experience with massage robots. AI can analyze the user's physical condition, health data, and personal preferences to design a personalized massage program. Through the integration of advanced intelligent sensors, massage robots can monitor users' physiological responses, such as muscle tension and body temperature, in real time. According to these data, massage robots can adjust their massage strength, speed, and focus area to overcome the "subhealth pain problem" of accurate positioning and efficient massage. At present, there are few studies on the clinical effectiveness of AI nudging robots, and patient self-reported changes in pain level, duration of pain relief, reduction in drug dependence, and objective measures of mobility (such as gait analysis) will be important indicators for evaluating their effectiveness in the future. When considering factors such as safety and comfort, AI data recording and analysis can also be used to measure the clinical efficacy of tuina robots.

AI-Directed Acupuncture Manipulation

Acupuncture, which is a therapeutic technique in TCM that has been practiced for thousands of years, has gained widespread global acceptance and demonstrated significant efficacy for various chronic diseases, particularly pain-related conditions. This therapeutic approach involves stimulating specific areas, known as acupoints, on the patient's body, thus eliciting sensations such as soreness, numbness, fullness, or heaviness, which is commonly referred to as "De Qi" or achieving qi [34]. Due to the inherent subjectivity and reliance on experience in traditional acupuncture practices, there is growing interest in parameter-based electroacupuncture to address these limitations [35]. By setting different parameters using an electroacupuncture device, clinical efficacy can be enhanced, thus facilitating further research. However, the efficacy of acupuncture is still not universally recognized [36], possibly for 2 main reasons. First, the inadequate design and implementation of past clinical research methods have led to a lack of clinical evidence. Second, the mechanism of acupuncture remains unclear, thus necessitating more high-quality evidence to elucidate its biological mechanisms for informed clinical decision-making [37]. The integration of AI and acupuncture shows great potential for substantially improving the precision of acupuncture prescriptions and treatment techniques. A bibliometric study by Zhou et al [38] demonstrated substantial progress in AI research within the acupuncture field over the past 2 decades, with significant contributions from the United States and China. However, the application of AI in acupuncture lacks a clear framework, with a scarcity of systematic research and a lack of organization of relevant technologies and application approaches.

Given the unique characteristics of AI and the importance of data mining in clinical acupuncture practice and manipulation, further research is needed [39]. Clinical trials are costly and limited, and most articles that analyze the safety and efficacy of acupuncture are of low quality and lack comprehensive analyses. There is still a lack of standardized acupuncture point selection protocols for many diseases [40]. Therefore, future efforts should focus on standardizing TCM while improving the quality of randomized controlled trials on acupuncture to obtain more and higher-quality clinical data, thus providing a foundation for AI-based clinical data mining. AI can analyze a patient's symptoms, signs, and physiological data and compare them to a large body of medical knowledge. Through machine learning and pattern recognition algorithms, AI can help clinicians interpret diagnostic data and provide potential pathological patterns or disease classifications that can help acupuncturists in developing treatment strategies. AI can then be used to analyze large amounts of clinical data and research the literature to determine the most effective point selection for a particular condition or disease situation. A recent study "linked" original studies and 332 systematic evaluations of evidence in 20 disease areas by using AI analysis techniques to

comprehensively improve clinical evidence for acupuncture therapy in the Epistemonikos database, which constructed a total of 77 evidence matrices [41]. This will facilitate the development of a machine learning framework to predict the efficacy of acupuncture and patient prognosis. Acupuncture manipulation techniques are crucial components of acupuncture therapy, and their efficacy is paramount [42]. However, the determination of the optimal stimulation intensity in clinical research is often challenging because of technique selection, treatment duration, needling speed, and force [43-45]. Therefore, the quantification and standardization of acupuncture manipulation, such as needle insertion force, duration, and direction, are essential for achieving clinical efficacy and AI-guided acupuncture manipulation. In response to the problems in standardizing operation techniques, AI technologies, especially machine learning models and sensor technologies, are being used to capture and analyze the nuances of manual needling operations. Acupuncture robots that are currently under development can accurately gauge the location of acupuncture points by measuring a person's height and sebum thickness and use ultrasound sensors to control the depth and speed of needling. These sensors and machine learning models are able to identify key parameters such as the needling force, speed, and angle to ensure the standardization and consistency of treatment. The application of sensor technology in acupuncture focuses on the precise control and measurement of the depth, force, and speed of needling. This robot uses an ultrasound sensor to control the depth and speed of needling. By emitting ultrasonic waves and receiving their echoes, the ultrasonic sensor can accurately measure the distance between the tip of the needle and the surface of the tissue to ensure that the depth of the needles is appropriate and avoid unnecessary injury to the patient. Through built-in mechanical sensors, the robot can also automatically adjust the needle insertion process according to changes in needle insertion resistance to ensure safe needle insertion.

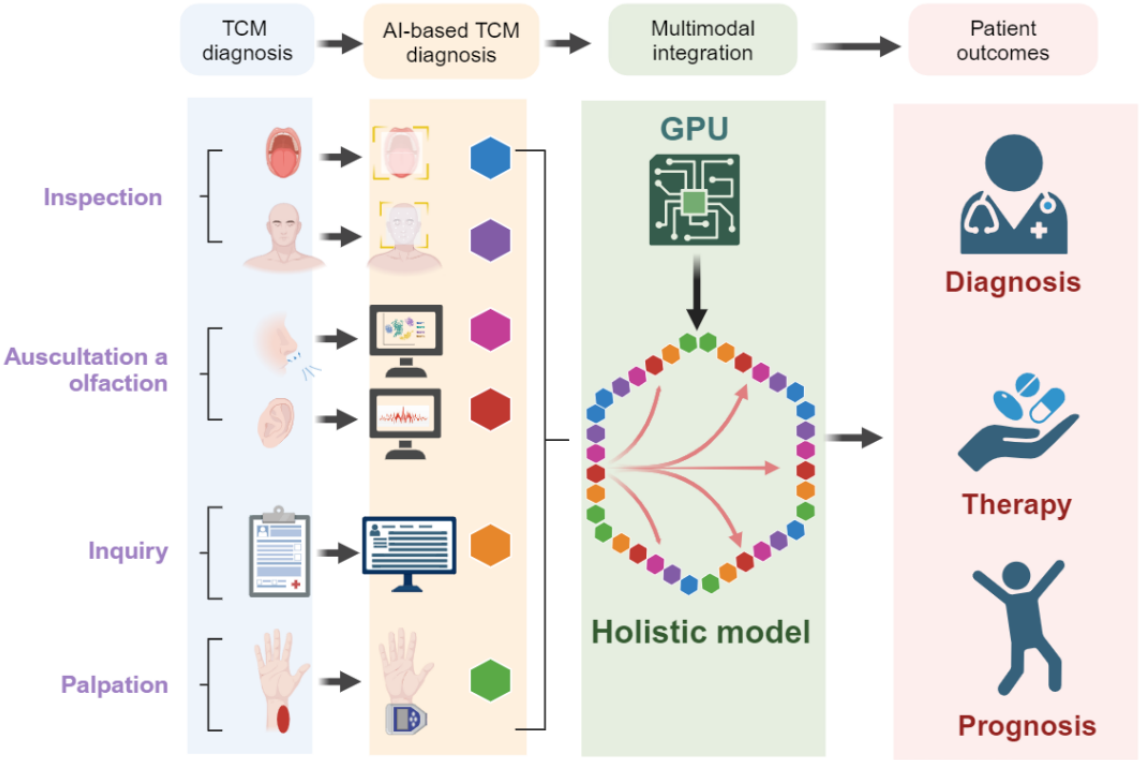
The standardization of acupuncture manipulation forms the basis of the use of AI in acupuncture. With AI technology, we propose three different ways to help standardize acupuncture: (1) imaging recognition–based standardization of acupuncture practitioners' techniques, (2) analysis of parameters derived from acupuncture practitioners' lifting and thrusting techniques using neural network image analysis systems, and (3) extraction of spatiotemporal features from video images of acupuncture operations by using computer vision technology [46]. In addition, a hybrid model that combines 3D convolutional neural networks and neural networks is used to recognize and classify dynamic hand gestures in acupuncture operation videos, thus enabling quantitative analyses and technical inheritance research for various techniques. Another approach involves recording acupuncture practitioners' movements and mechanical parameters during acupuncture procedures by using 3-axis posture sensors. Davis et al [47] developed force and motion sensor technology (acusensors) to quantify the linear and rotational movements of acupuncture needles and the force and torque that are generated during manual needle manipulation. A standardized TCM acupuncture manipulation database was established for the quantification of motion and force patterns. These data serve as a crucial tool for future AI applications in acupuncture. Finally, acupuncture parameters based on other electrophysiological signals have been recorded, thus showing significant differences in electrophysiological signals between acupuncture points and nearby nonacupuncture points and highlighting the electrical specificity of acupoints [48,49]. This finding serves as compelling evidence for TCM theory and provides parameters for the standardization of acupuncture stimulation. In addition, collaboration between AI experts, acupuncturists, and biomedical engineers is essential for developing and improving acupuncture-related technologies. The data analysis and intelligent algorithms that are provided by AI experts can help acupuncturists better understand treatment effects and optimize treatment plans. The clinical experience and theoretical knowledge of acupuncturists can guide AI experts in developing intelligent systems that better meet clinical needs. Moreover, there are some prominent conditions or events existing outside of normal circumstances that exist beyond the abilities of AI. In such cases, acupuncturists can make judgments based on their own experience and knowledge. Technical support from biomedical engineers subsequently ensures that these intelligent systems can be effectively applied in practice. Close collaboration among the 3 factors is the key to promoting technological innovation in acupuncture, improving treatment outcomes, and standardizing and popularizing acupuncture techniques. Through this interdisciplinary collaboration, modern technology can be better used to enhance the value and impact of traditional acupuncture medicine. AI technology can be used to simulate acupuncture operations and provide support for learning and training. For example, through virtual reality and augmented reality technologies, AI can create simulated acupuncture treatment scenarios that allow learners to practice acupuncture techniques and decision-making processes in a virtual environment. This approach improves learning efficiency and reduces risks in actual practice.

Outlook

The future is promising for the integration of AI into TCM diagnosis (Figure 1). The core of TCM diagnosis involves 4 fundamental methods: inspection, auscultation and olfaction, inquiry, and palpation. First, through inspection, we can explore the application of AI in facial and tongue diagnosis. By leveraging image processing techniques, we anticipate groundbreaking innovations in these areas. Due to the low contrast and noisy nature of ultrasound images, which require extensive knowledge of tongue structure and ultrasound data interpretation, it can be challenging for medical professionals to accurately diagnose various conditions. To overcome these challenges, researchers have proposed new deep neural networks, referred to as BowNet models [50]. These models leverage the global prediction capabilities of encoder-decoder fully convolutional neural networks and the high-resolution extraction features of dilated convolutions. The models are designed to automatically extract and track the tongue contour in real-time ultrasound data. By providing more accurate and objective tongue contour tracking, AI can assist TCM practitioners in making more informed diagnoses and treatment decisions. However, the digitization of valuable empirical data

from TCM facial and tongue diagnosis remains a challenge, although unsupervised machine learning may provide a solution. First, addressing the challenge of digitizing TCM experience data often requires seamless collaboration among diverse teams, which encompasses TCM practitioners, AI experts, and health care institutions. This initiative involves partnering with medical facilities to meticulously gather and curate a rich array of TCM clinical case data, thus encompassing detailed patient complaints, symptomatic expressions, tongue diagnoses, and pulse assessments. Subsequently, synergy between TCM professionals and AI researchers has extended to the intricate task of knowledge structuring. The translation of the profound insights of TCM theories, herbal formulas, and medicinal properties into a web-based knowledge graph can greatly enhance the accessibility and utility of this ancient medical wisdom. This graph facilitates a nuanced understanding of intricate relationships within TCM, underpins the development of intelligent recommendation systems, and assists in informed decision-making within the realm of TCM. Through these collaborative efforts, the rich tapestry of TCM experience data is meticulously incorporated into accessible, digital formats, thereby bridging the gap between ancient wisdom and modern technological advancements. These initiatives have propelled the modernization and globalization of TCM and significantly contributed to the broader landscape of health care innovation. Second, TCM diagnosis through auscultation and olfaction provides new avenues for development. We can develop specialized AI tools for auditory and olfactory systems that can complement TCM diagnosis. Third, the analysis of inquiry in TCM diagnosis presents an exciting frontier. There has been significant progress in Western medicine in this area with regard to LLMs, and TCM EHRs contain unique knowledge graphs that are specific to TCM diagnosis. Therefore, fine-tuning is necessary for the analysis of TCM EHRs. Fourth, pulse diagnosis is a cornerstone of TCM diagnosis and requires the development of more advanced pulse detection tools. By obtaining substantial pulse data, we can standardize objective pulse analysis. By leveraging machine learning and classification techniques, we can align these data with TCM pulse patterns, thus ultimately achieving AI-driven pulse diagnosis. Quantifiable metrics or benchmarks are instrumental in ensuring the high performance of AI systems, establishing unambiguous objectives, and measuring advancements in the integration of AI technologies. For example, metrics such as diagnostic accuracy and generalization capability can assess AI systems' proficiency in managing patient data across varying regions, age groups, and sexes, thereby ensuring their efficacy across a wide array of populations. By juxtaposing outcomes that are generated by AI systems with diagnoses that are rendered by seasoned TCM experts, the precision of AI systems in pinpointing specific conditions can be quantified. This quantification is facilitated through the computation of statistical indicators such as sensitivity (true positive rate), specificity (true negative rate), precision, and the F_1-score. These benchmarks serve as a foundation for the continuous refinement of AI systems, thus ensuring their optimal integration and application within the sphere of TCM. These directions of development have the potential to revolutionize TCM diagnosis, thus enhancing its accuracy and efficiency.

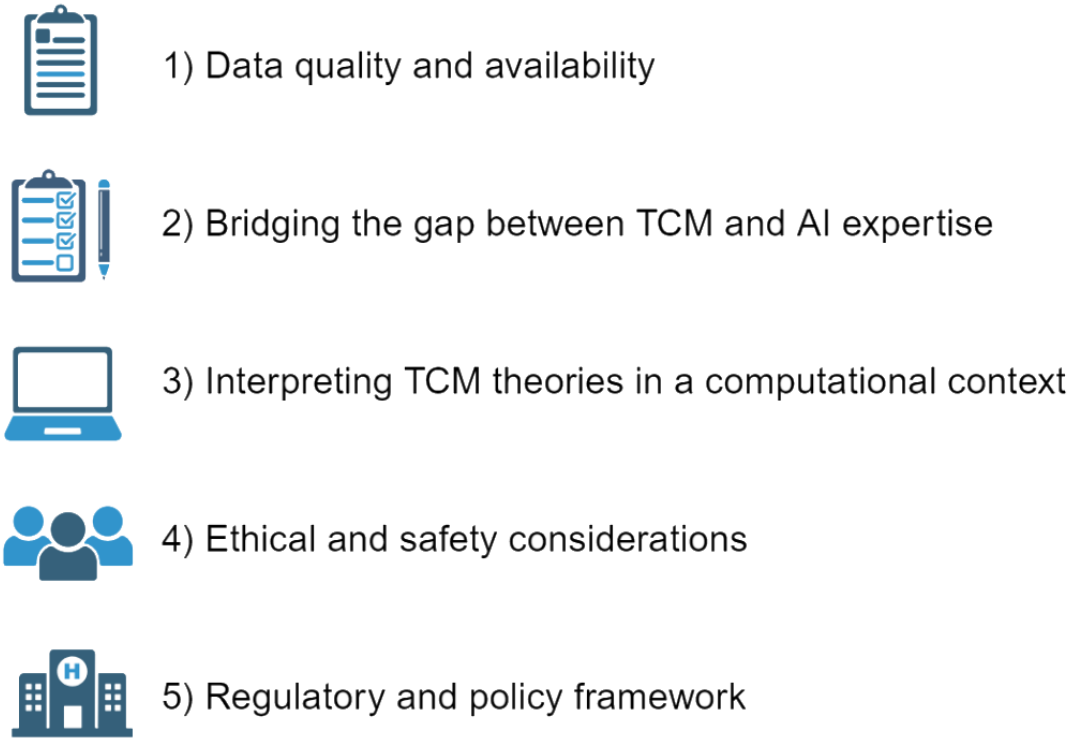
Figure 1. Overview of artificial intelligence (AI) development strategies based on traditional Chinese medicine (TCM) diagnosis. The acquisition and standardization of unimodal data through TCM diagnostic techniques is followed by the integration of multimodal data using a comprehensive model. This approach aids in enhancing predictions and supports TCM diagnoses for treatment and prognosis. GPU: graphics processing unit.



Challenges

There are unique challenges to the use of AI in TCM (Figure 2). First, regarding data quality and availability, the successful implementation of AI in TCM relies on access to reliable and standardized data sets. High-quality data can potentially promote clinical diagnosis and treatment in precision TCM. However, data collection and digitization efforts in TCM can be challenging, and the quality and interoperability of existing data sets may vary. Varied interpretations of identical conditions among TCM practitioners can result in divergent diagnostic criteria and terminological applications. Such disparities, coupled with the potential for errors and biases in manually entered data, can significantly impact the learning efficacy of AI systems. Consequently, the establishment of standardized TCM diagnostic criteria and terminology glossaries, in addition to the implementation of uniform data entry protocols for TCM practitioners, is essential for mitigating interpractitioner discrepancies. Furthermore, the use of advanced automated data collection techniques, including image recognition and natural language processing, is instrumental in enhancing the quality and precision of the collected data. These measures collectively contribute to the refinement of the analytical capabilities of AI within the realm of TCM, thus ensuring a more accurate and reliable diagnostic process. Second, to bridge the gap between TCM and AI expertise, the integration of AI technologies into TCM requires collaboration and communication among TCM practitioners and AI experts. Bridging the gap between these domains is crucial for developing AI algorithms that align with TCM principles and meet specific clinical needs. Measures could be taken to promote collaboration between TCM organizations and technology companies, as well as higher-education institutions, to facilitate the development of AI-driven TCM diagnostic and therapeutic tools. These collaborations will foster innovation and create platforms for TCM students and practitioners to perform internships in the technology industry. In addition, the provision of scholarships and research grants is critical for incentivizing and sustaining interdisciplinary scholarship. By allowing students and researchers to delve into the convergence of TCM and AI, we can accelerate the digitization of TCM knowledge. Third, TCM theories must be interpreted in a computational context. TCM theories are often complex and based on holistic and individualized perspectives. The translation of this knowledge into AI algorithms and computational models is a significant challenge that requires careful consideration of cultural, philosophical, and theoretical aspects. Many concepts in Chinese medicine, such as qi, yin and yang, and the 5 elements, are abstract and ambiguous, and it is difficult to describe these concepts in precise mathematical language; therefore, ambiguous logic can be used to address ambiguous concepts in TCM, and Bayesian networks can be used to simulate causality and uncertainty in the theory of Chinese medicine. Fourth, there are notable ethical and safety considerations regarding this scenario, as with any implementation of AI in health care [51]. Ensuring patient privacy, data security, and transparency in algorithm decision-making is essential for building trust and ethical practices in AI-supported TCM. Informed patient consent must be obtained before collecting and using patient data. This includes a full explanation of the purpose of data collection, how the data will be used, how long they will be stored, and potential risks. To protect patients' privacy, all the data sets that are used for machine learning should be anonymized by removing or encrypting any personally identifiable information. During the development and deployment of AI systems, ethical review committees need to be established to ethically review the design, implementation, and evaluation of AI systems to ensure that all the activities meet ethical standards. Fifth, the integration of AI into TCM necessitates clear regulatory frameworks and policies that govern its implementation, including issues related to data protection, algorithm validation, and clinical decision-making. Sixth, the function and efficacy of TCM are broadly accepted worldwide; however, the underlying mechanism has remained enigmatic, thus limiting people's confidence in TCM and precision therapy that is learned by AI. In summary, the process of implementing and validating AI tools in a clinical setting requires careful planning and rigorous execution. Representative and actionable clinical environments are selected to develop pilot projects. These projects should focus on specific TCM diagnostic tasks, such as tongue analysis, pulse recognition, and symptom assessment. Clinical trials need to be designed and executed to evaluate the performance of AI tools in real clinical settings. This includes randomized controlled trials and prospective cohort studies to assess the impact of AI tools on patient outcomes. Finally, the results and experiences will be published and shared through academic journals and conferences to promote communication and learning within the industry.

Figure 2. Summary of the challenges of integrating artificial intelligence (AI) into traditional Chinese medicine (TCM) diagnosis.



Conclusions

In conclusion, the integration of AI into TCM exhibits immense promise for improving diagnosis, including inspection, auscultation and olfaction, inquiry, and palpation. The successful integration of AI into TCM is evident through advancements in areas such as image analysis for tongue diagnosis, the development of intelligent tuina massage systems, and the application of machine learning to refine treatment protocols based on individual patient data. Addressing the challenges of data quality, the standardization of data sets, interdisciplinary collaboration, the interpretation of TCM theories, ethical considerations, and regulatory frameworks is crucial for the successful and responsible implementation of AI in TCM. By overcoming these challenges, we can leverage the power of AI to enhance patient care, personalize treatments, and advance our understanding of TCM. Moreover, we can develop more precise AI models that are tailored to TCM, thus creating a positive cycle of problem-solving and progress that ultimately leads to better patient care. By combining the wisdom of TCM with the power of AI technology, we can improve patient outcomes and promote the integration of TCM into modern health care systems. It is imperative to conduct more research into AI's ability to decode complex diagnostic patterns that are inherent to TCM. The validation of AI-enhanced TCM treatment methods through clinical trials is essential to ensure their safety and efficacy, thus providing empirical support for their widespread adoption. As we advance the integration of AI into TCM, it is vital to uphold ethical standards that prioritize patient rights, cultural integrity, and data privacy. The responsible use of AI will ensure that technological advancements align with the principles and practices of TCM, thus safeguarding the well-being of patients and respecting the cultural significance of this ancient medical system. The fusion of AI with TCM has the potential to bridge traditional and modern medical practices, enrich global health, and foster cultural exchange. By integrating these 2 domains, we can create a more comprehensive health care system that is both innovative and respectful of historical practices.

Finally, a call to action is made to all stakeholders (practitioners, researchers, policy makers, and investors) to collaborate and support the integration of AI into TCM. Through collective efforts, we can harness AI to transform patient care, broaden our understanding of TCM on a global scale, and identify new horizons in health care that are both deeply rooted in tradition and boldly futuristic.

Acknowledgments

This research was funded by grants from the Ministry of Science and Technology of China and the National Key R&D Program of China (2023YFC2506802). The images in Figures 1 and 2 were created using BioRender [52].

Authors' Contributions

All the authors participated in the conceptualization, methodology, validation, and writing of the manuscript.

Conflicts of Interest

None declared.

References

1. Wang Y, Shi X, Li L, Efferth T, Shang D. The impact of artificial intelligence on traditional Chinese medicine. Am J Chin Med 2021;49(6):1297-1314. [doi: 10.1142/S0192415X21500622] [Medline: 34247564]
2. Zhang S, Wang W, Pi X, He Z, Liu H. Advances in the application of traditional Chinese medicine using artificial intelligence: a review. Am J Chin Med 2023;51(5):1067-1083. [doi: 10.1142/S0192415X23500490] [Medline: 37417927]
3. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 2019 Mar;69(2):127-157 [FREE Full text] [doi: 10.3322/caac.21552] [Medline: 30720861]
4. Bondi E, Maggioni E, Brambilla P, Delvecchio G. A systematic review on the potential use of machine learning to classify major depressive disorder from healthy controls using resting state fMRI measures. Neurosci Biobehav Rev 2023 Jan;144:104972. [doi: 10.1016/j.neubiorev.2022.104972] [Medline: 36436736]
5. Sun J, Dong QX, Wang SW, Zheng YB, Liu XX, Lu TS, et al. Artificial intelligence in psychiatry research, diagnosis, and therapy. Asian J Psychiatr 2023 Sep;87:103705. [doi: 10.1016/j.ajp.2023.103705] [Medline: 37506575]
6. Wang Z, Wang D, Liu W, Wang Z. Traditional Chinese medicine diagnosis and treatment based on systematics. iLIVER 2023 Dec;2(4):181-187. [doi: 10.1016/j.iliver.2023.08.004]
7. Maciocia G. Diagnosis in Chinese Medicine: A Comprehensive Guide. Amsterdam, The Netherlands: Churchill Livingstone; 2004.
8. Zhang YH, Lv J, Gao W, Li J, Fang JQ, He LY, et al. Practitioners' perspectives on evaluating treatment outcomes in traditional Chinese medicine. BMC Complement Altern Med 2017 May 18;17(1):269 [FREE Full text] [doi: 10.1186/s12906-017-1746-8] [Medline: 28521826]
9. Lu LM, Chen X, Xu JT. Determination methods for inspection of the complexion in traditional Chinese medicine: a review [Article in Chinese]. Zhong Xi Yi Jie He Xue Bao 2009 Aug;7(8):701-705. [doi: 10.3736/jcim20090801] [Medline: 19671406]
10. Huang Z, Miao J, Chen J, Zhong Y, Yang S, Ma Y, et al. A traditional Chinese medicine syndrome classification model based on cross-feature generation by convolution neural network: model development and validation. JMIR Med Inform 2022 Apr 06;10(4):e29290 [FREE Full text] [doi: 10.2196/29290] [Medline: 35384854]
11. Liu Q, Li Y, Yang P, Liu Q, Wang C, Chen K, et al. A survey of artificial intelligence in tongue image for disease diagnosis and syndrome differentiation. Digit Health 2023 Aug 06;9:20552076231191044 [FREE Full text] [doi: 10.1177/20552076231191044] [Medline: 37559828]
12. Chiu CC, Chang HH, Yang CH. Objective auscultation for traditional Chinese medical diagnosis using novel acoustic parameters. Comput Methods Programs Biomed 2000 Jun;62(2):99-107. [doi: 10.1016/s0169-2607(00)00055-9] [Medline: 10764936]
13. Lee BK, Mayhew EJ, Sanchez-Lengeling B, Wei JN, Qian WW, Little KA, et al. A principal odor map unifies diverse tasks in olfactory perception. Science 2023 Sep;381(6661):999-1006. [doi: 10.1126/science.ade4401] [Medline: 37651511]
14. Fu W, Xu L, Yu Q, Fang J, Zhao G, Li Y, et al. Artificial intelligent olfactory system for the diagnosis of Parkinson's disease. ACS Omega 2022 Jan 26;7(5):4001-4010 [FREE Full text] [doi: 10.1021/acsomega.1c05060] [Medline: 35155895]
15. Li M, Wen G, Zhong J, Yang P. Personalized intelligent syndrome differentiation guided by TCM consultation philosophy. J Healthc Eng 2022 Nov 07;2022:6553017 [FREE Full text] [doi: 10.1155/2022/6553017] [Medline: 36389107]
16. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med 2022 Dec 26;5(1):194 [FREE Full text] [doi: 10.1038/s41746-022-00742-2] [Medline: 36572766]
17. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep 2023 Nov 22;13(1):20512 [FREE Full text] [doi: 10.1038/s41598-023-46995-z] [Medline: 37993519]
18. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
19. Wang YY, Wang SH, Jan MY, Wang WK. Past, present, and future of the pulse examination (mài zhěn). J Tradit Complement Med 2012 Jul;2(3):164-185 [FREE Full text] [doi: 10.1016/s2225-4110(16)30096-7] [Medline: 24716130]
20. Lan KC, Litscher G, Hung TH. Traditional Chinese medicine pulse diagnosis on a smartphone using skin impedance at acupoints: a feasibility study. Sensors (Basel) 2020 Aug 17;20(16):4618 [FREE Full text] [doi: 10.3390/s20164618] [Medline: 32824477]
21. Liu Z, Zhang L, Wu J, Zheng Z, Gao J, Lin Y, et al. Machine learning-based classification of circadian rhythm characteristics for mild cognitive impairment in the elderly. Front Public Health 2022 Oct 28;10:1036886 [FREE Full text] [doi: 10.3389/fpubh.2022.1036886] [Medline: 36388285]

22. Yao Y, Zhou S, Alastruey J, Hao L, Greenwald SE, Zhang Y, et al. Estimation of central pulse wave velocity from radial pulse wave analysis. Comput Methods Programs Biomed 2022 Jun;219:106781. [doi: 10.1016/j.cmpb.2022.106781] [Medline: 35378395]
23. Liu Y, Bai X, Zhang H, Zhi X, Jiao J, Wang Q, et al. Efficacy and safety of tuina for senile insomnia: a protocol for systematic review and meta-analysis. Medicine (Baltimore) 2022 Feb 25;101(8):e28900 [FREE Full text] [doi: 10.1097/MD.0000000000028900] [Medline: 35212294]
24. Zhu Q, Li J, Fang M, Gong L, Sun W, Zhou N. [Effect of Chinese massage (Tui Na) on isokinetic muscle strength in patients with knee osteoarthritis]. J Tradit Chin Med 2016 Jun;36(3):314-320 [FREE Full text] [doi: 10.1016/s0254-6272(16)30043-7] [Medline: 27468545]
25. Cheng ZJ, Zhang SP, Gu YJ, Chen ZY, Xie FF, Guan C, et al. Effectiveness of Tuina therapy combined with Yijinjing exercise in the treatment of nonspecific chronic neck pain: a randomized clinical trial. JAMA Netw Open 2022 Dec 01;5(12):e2246538 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.46538] [Medline: 36512354]
26. Liu D, Zhang YQ, Yu TY, Han SL, Xu YJ, Guan Q, et al. Regulatory mechanism of the six-method massage antipyretic process on lipopolysaccharide-induced fever in juvenile rabbits: a targeted metabolomics approach. Heliyon 2023 Dec 08;10(1):e23313 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e23313] [Medline: 38148795]
27. Lei LM, Wu GF, Qiu SY. Literature analysis of tuina in regulating subhealth status in recent 10 years. J Acupunct Tuina Sci 2014 Feb 1;12:60-66. [doi: 10.1007/s11726-014-0749-y]
28. Liu Y, Cao L, Liu J, Zhang Z, Fan P, Zhu Y, et al. Increased hippocampal glucocorticoid receptor expression and reduced anxiety-like behavior following Tuina in a rat model with allergic airway inflammation. J Manipulative Physiol Ther 2022 Oct;45(8):586-594. [doi: 10.1016/j.jmpt.2023.04.008] [Medline: 37294215]
29. Zhang Y, Zhang W. Recent patents on Chinese massage robot. Recent Pat Eng 2017 Dec 01;11(3):156-161. [doi: 10.2174/2212797610666170127171713]
30. Pang Z, Zhang B, Yu J, Sun Z, Gong L. Design and analysis of a Chinese medicine based humanoid robotic arm massage system. Appl Sci 2019 Oct 12;9(20):4294. [doi: 10.3390/app9204294]
31. Yang J, Croghan IT, Fokken SC, Johnson DE, Calva JJ, Do A, et al. Satisfaction and feasibility evaluation of an electronic massager-expert manipulative massage automation (EMMA): a pilot study. J Prim Care Community Health 2023;14:21501319231199010 [FREE Full text] [doi: 10.1177/21501319231199010] [Medline: 37698255]
32. Johnson JA. FDA regulation of medical devices. Congressional Research Service. 2016 Sep 14. URL: https://crsreports.congress.gov/product/pdf/R/R42130 [accessed 2024-06-19]
33. Sijia L. New law sparks the expectation over the future of traditional Chinese medicine: can TCM law effectively promote the development of TCM industry in China? Med Law 2018 Mar;37(1):193-228 [FREE Full text]
34. Chen T, Zhang WW, Chu YX, Wang YQ. Acupuncture for pain management: molecular mechanisms of action. Am J Chin Med 2020;48(4):793-811. [doi: 10.1142/S0192415X20500408] [Medline: 32420752]
35. Zhang YY, Chen QL, Wang Q, Ding SS, Li SN, Chen SJ, et al. Role of parameter setting in electroacupuncture: current scenario and future prospects. Chin J Integr Med 2022 Oct;28(10):953-960. [doi: 10.1007/s11655-020-3269-2] [Medline: 32691284]
36. Yang C, Hao Z, Zhang LL, Guo Q. Efficacy and safety of acupuncture in children: an overview of systematic reviews. Pediatr Res 2015 Aug;78(2):112-119. [doi: 10.1038/pr.2015.91] [Medline: 25950453]
37. Ee C, Xue C, Chondros P, Myers SP, French SD, Teede H, et al. Acupuncture for menopausal hot flashes: a randomized trial. Ann Intern Med 2016 Feb 02;164(3):146-154. [doi: 10.7326/M15-1380] [Medline: 26784863]
38. Zhou Q, Zhao T, Feng K, Gong R, Wang Y, Yang H. Artificial intelligence in acupuncture: a bibliometric study. Math Biosci Eng 2023 Apr 27;20(6):11367-11378 [FREE Full text] [doi: 10.3934/mbe.2023504] [Medline: 37322986]
39. Zhao S, Huang T. Application of artificial intelligence in acupuncture and moxibustion. Int J Clin Acupunct 2022;31(3):224. [doi: 10.3103/S1047197922030061]
40. Peixun Y, Bing G, Yujun X. Meta-analysis of the effect of distal or local point selection on acupuncture efficacy. World J Acupunct Moxibustion 2018 Jun;28(2):114-120. [doi: 10.1016/j.wjam.2018.05.005]
41. Lu L, Zhang Y, Tang X, Ge S, Wen H, Zeng J, et al. Evidence on acupuncture therapies is underused in clinical practice and health policy. BMJ 2022 Feb 25;376:e067475 [FREE Full text] [doi: 10.1136/bmj-2021-067475] [Medline: 35217525]
42. Stux G, Berman B, Pomeranz B. Basics of Acupuncture, Fifth Edition. Berlin, Heidelberg: Springer; 2003.
43. Xing W, Li Q. Effects of different manipulations of acupuncture on electrical activity of stomach in humans. J Tradit Chin Med 1998 Mar;18(1):39-42. [Medline: 10437261]
44. Huang T, Huang X, Zhang W, Jia S, Cheng X, Litscher G. The influence of different acupuncture manipulations on the skin temperature of an acupoint. Evid Based Complement Alternat Med 2013;2013:905852 [FREE Full text] [doi: 10.1155/2013/905852] [Medline: 23476709]
45. Tang W, Yang H, Liu T, Gao M, Xu G. Study on quantification and classification of acupuncture lifting-thrusting manipulations on the basis of motion video and self-organizing feature map neural network. Shanghai J Acupunct Moxibustion 2017(12):1012-1020.

46. Zhu M, Liu DM, Pei J, Zhan YJ, Shen HY. An acupuncture manipulation classification system based on three-axis attitude sensor and computer vision. Zhen Ci Yan Jiu 2023 Dec 25;48(12):1274-1281. [doi: 10.13702/j.1000-0607.20221145] [Medline: 38146251]
47. Davis RT, Churchill DL, Badger GJ, Dunn J, Langevin HM. A new method for quantifying the needling component of acupuncture treatments. Acupunct Med 2012 Jun;30(2):113-119 [FREE Full text] [doi: 10.1136/acupmed-2011-010111] [Medline: 22427464]
48. Hamed Mozaffari M, Lee WS. Encoder-decoder CNN models for automatic tracking of tongue contours in real-time ultrasound data. Methods 2020 Jul 01;179:26-36. [doi: 10.1016/j.ymeth.2020.05.011] [Medline: 32450205]
49. Zhou Q, Gai S, Lin N, Zhang J, Zhang L, Yu R, et al. Power spectral differences of electrophysiological signals detected at acupuncture points and non-acupuncture points. Acupunct Electrother Res 2014;39(2):169-181. [doi: 10.3727/036012914x14054537750508] [Medline: 25219030]
50. Kim M, Seo HD, Sawada K, Ishida M. Study of biosignal response during acupuncture points stimulations. In: Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2008 Presented at: IEMBS 2008; August 20-25, 2008; Vancouver, BC. [doi: 10.1109/iembs.2008.4649246]
51. Feng C, Zhou S, Qu Y, Wang Q, Bao S, Li Y, et al. Overview of artificial intelligence applications in Chinese medicine therapy. Evid Based Complement Alternat Med 2021 Mar 17;2021:6678958 [FREE Full text] [doi: 10.1155/2021/6678958] [Medline: 33815559]
52. BioRender. URL: https://www.biorender.com/ [accessed 2024-06-24]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
EMMA: Expert Manipulative Massage Automation
LLM: large language model
TCM: traditional Chinese medicine

Edited by G Eysenbach, A Castonguay; submitted 18.03.24; peer-reviewed by J Sun, X Zhang; comments to author 05.04.24; revised version received 10.05.24; accepted 31.05.24; published 28.06.24.

Please cite as:
Lu L, Lu T, Tian C, Zhang X
AI: Bridging Ancient Wisdom and Modern Innovation in Traditional Chinese Medicine
JMIR Med Inform 2024;12:e58491
URL: https://medinform.jmir.org/2024/1/e58491
doi:10.2196/58491
PMID:

Viewpoint

Considerations for Quality Control Monitoring of Machine Learning Models in Clinical Practice

Louis Faust[1], PhD; Patrick Wilson[1], MPH; Shusaku Asai[1], MS; Sunyang Fu[2], PhD; Hongfang Liu[2], PhD; Xiaoyang Ruan[2], PhD; Curt Storlie[1], PhD

[1]Robert D and Patricia E Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, United States

[2]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Louis Faust, PhD
Robert D and Patricia E Kern Center for the Science of Health Care Delivery
Mayo Clinic
Mayo Clinic, 200 First St. SW
Rochester, MN, 55905
United States
Phone: 1 (507) 284 2511
Email: Faust.Louis@mayo.edu

Abstract

Integrating machine learning (ML) models into clinical practice presents a challenge of maintaining their efficacy over time. While existing literature offers valuable strategies for detecting declining model performance, there is a need to document the broader challenges and solutions associated with the real-world development and integration of model monitoring solutions. This work details the development and use of a platform for monitoring the performance of a production-level ML model operating in Mayo Clinic. In this paper, we aimed to provide a series of considerations and guidelines necessary for integrating such a platform into a team's technical infrastructure and workflow. We have documented our experiences with this integration process, discussed the broader challenges encountered with real-world implementation and maintenance, and included the source code for the platform. Our monitoring platform was built as an R shiny application, developed and implemented over the course of 6 months. The platform has been used and maintained for 2 years and is still in use as of July 2023. The considerations necessary for the implementation of the monitoring platform center around 4 pillars: feasibility (what resources can be used for platform development?); design (through what statistics or models will the model be monitored, and how will these results be efficiently displayed to the end user?); implementation (how will this platform be built, and where will it exist within the IT ecosystem?); and policy (based on monitoring feedback, when and what actions will be taken to fix problems, and how will these problems be translated to clinical staff?). While much of the literature surrounding ML performance monitoring emphasizes methodological approaches for capturing changes in performance, there remains a battery of other challenges and considerations that must be addressed for successful real-world implementation.

(JMIR Med Inform 2024;12:e50437) doi:10.2196/50437

KEYWORDS

artificial intelligence; machine learning; implementation science; quality control; monitoring; patient safety

Introduction

As machine learning (ML) models integrate into clinical practice, ensuring their continued efficacy becomes a critical task. A pervasive limitation in ML is the inability of most models to adapt to changes in their environment over time. As a result, a model that may have performed exceptionally in its development environment can become gradually or immediately less accurate while in production [1,2]. This problem has been well studied by the ML community, with current literature offering invaluable methodological strategies for the detection of declining model performance and the ethical implications of such declines [3-7]. However, the proper choice of monitoring algorithm is only one step in the larger series of problems and considerations surrounding the sustained maintenance of these models in a real-world scenario. While some authors address the wider set of problems encountered in the long-term maintenance strategy of a deployed model, it is typically only an acknowledgment of these problems, rather than the personal experiences and solutions developed to solve them [8,9]. As

such, we aimed to supplement current literature with an alternative approach in which we provided an in-depth review of the experiences and challenges encountered when integrating our ML monitoring solution into clinical practice.

This paper focuses on an ML model implemented into Mayo Clinic's practice in 2018. The model, known as "Control Tower," is a fully integrated health care delivery model that predicts the need for inpatient palliative care through modeling palliative care consultation. The model runs automatically on all inpatients at Mayo Clinic's St Marys and Methodist Hospitals in Rochester, Minnesota, with patient scores monitored by the palliative care practice [10]. The approach was to treat the palliative care consult as a time-to-event outcome. Some of the features used are static (patient demographics and prior history), while others are time varying or dynamic (such as laboratory values, vitals, and in-hospital events). To capture the time-varying nature of these covariates, we used a heterogeneous Poisson process. Furthermore, it was crucial to account for nonlinearity and interactions; as a result, we used a gradient boosting machine. The model was validated through a clinical trial conducted from 2019 to 2022 to assess real-world effectiveness and is still in use by the palliative care practice as of July 2023 [11,12]. The study by Murphree et al [10] provides a complete methodological overview of the ML model and validation procedure. The Control Tower monitoring platform was developed and implemented over the course of 6 months. The platform has been used and maintained for 2 years and is still in use as of July 2023.

This paper provides a series of guidelines for developing and integrating ML performance monitoring into a team's workflow. Guidelines were developed from real-world experiences and challenges encountered throughout this process by a data science team at Mayo Clinic. In addition, a comprehensive overview of the developed monitoring platform is provided, as well as the accompanying source code for demonstration purposes (Multimedia Appendix 1). Overall, this paper serves as a primer for considerations that must be made when implementing and maintaining a model-monitoring system in a clinical setting, coupled with the corresponding solutions that our team had used.

Development of the Model Monitoring Platform

Overview

Traditionally, guidelines are developed through expert-driven processes, such as the Delphi method that seeks to provide standards through initial conceptions followed by several rounds of revisions until ultimately converging to an agreed-upon set [13]. However, in emerging areas where expertise is sparse, expert-driven approaches are often costly when seeking consensus of multiple experts through multiple rounds of responses [14]. An alternative to the expert-driven approach is experience-driven methodologies, which emphasize the personal experiences and observations of individuals who have directly encountered the phenomena. Normally these methodologies focus on practical knowledge through the explication of the "real world." Our team opted to derive a set of guidelines based on our specific real-world experiences and the challenges faced when designing, implementing, and integrating the Control Tower monitoring platform. Our specific methodologies used throughout this process are documented here and later generalized into a series of guidelines in the *Design Considerations* section.

Establishing the Team and Responsibilities

When planning the phases of Control Tower, it was decided that the role of monitoring the model would remain with the model development team. The task of monitoring was divided among 4 team members, rotating the responsibility of monitoring, monthly. This approach ensured monitoring would not significantly inhibit the bandwidth of any 1 team member. Monitoring responsibilities did not fall to the team member who developed the model, as their primary task in monitoring would be to retrain the model when necessary. The monitoring platform was checked biweekly, Mondays and Thursdays, to balance coverage and analyst time. The Monday check ensured immediate response to any issues that may have occurred during the previous weekend, and the Thursday check provided enough time before the upcoming weekend to identify and resolve any errors that may have occurred during the week. Typically, a single-model monitoring session would take approximately 5 to 10 minutes, assuming no problems were encountered.

Platform Development

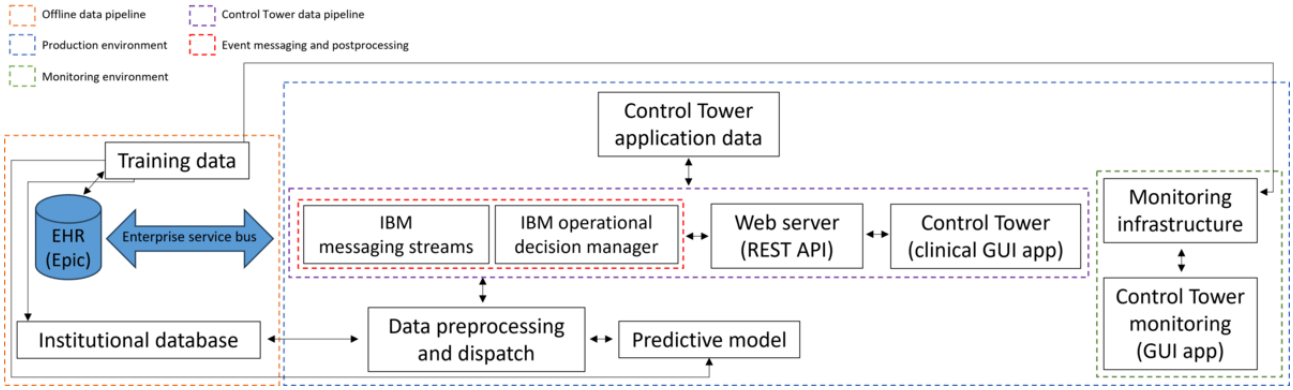
Overview

Performance monitoring of Control Tower was accomplished through the development of an R Shiny web application that comprised data visualizations and interactive tables. The goal was to create a centralized, user-friendly platform for all team members to check model performance. The platform consisted of 5 different tabs addressing different types of data shift, providing multiple degrees of granularity depending on the depth of investigation required. The model used a set of 126 features, measured daily, and was called an average of 80,000 times per day. Daily metrics collected for performance monitoring included mean and scale covariate shifts per feature, predicted probabilities, and the number of daily predictions made by the model. The resulting data size of these collected performance monitoring metrics was trivial; however, capturing patient-generated data resulted in data creation on the order of GB per day, requiring a dedicated storage space.

Figure 1 provides an overview of the system architecture for the Control Tower model and monitoring platform. The figure details the offline data pipeline used for the initial training of the Control Tower model; the components of the broader production environment and pipelines necessary for the predictive model and clinical graphical user interface (GUI) app; and finally, the components necessary for monitoring the performance of the Control Tower model. A more detailed visualization and comprehensive description of the system architecture is provided by Murphree et al [10]. Briefly, they outlined our deployment strategy which integrates a Representational State Transfer application programming interface within a Docker container, enabling the integration of

predictive models into the Control Tower GUI. The data ingestion and preprocessing pipeline, integrated with IBM Streams and Operational Decision Manager, facilitates real-time prediction processing triggered by updates to institutional health records (Health Level Seven messages by our electronic health record). The Control Tower GUI application is built with Angular (Google LLC).

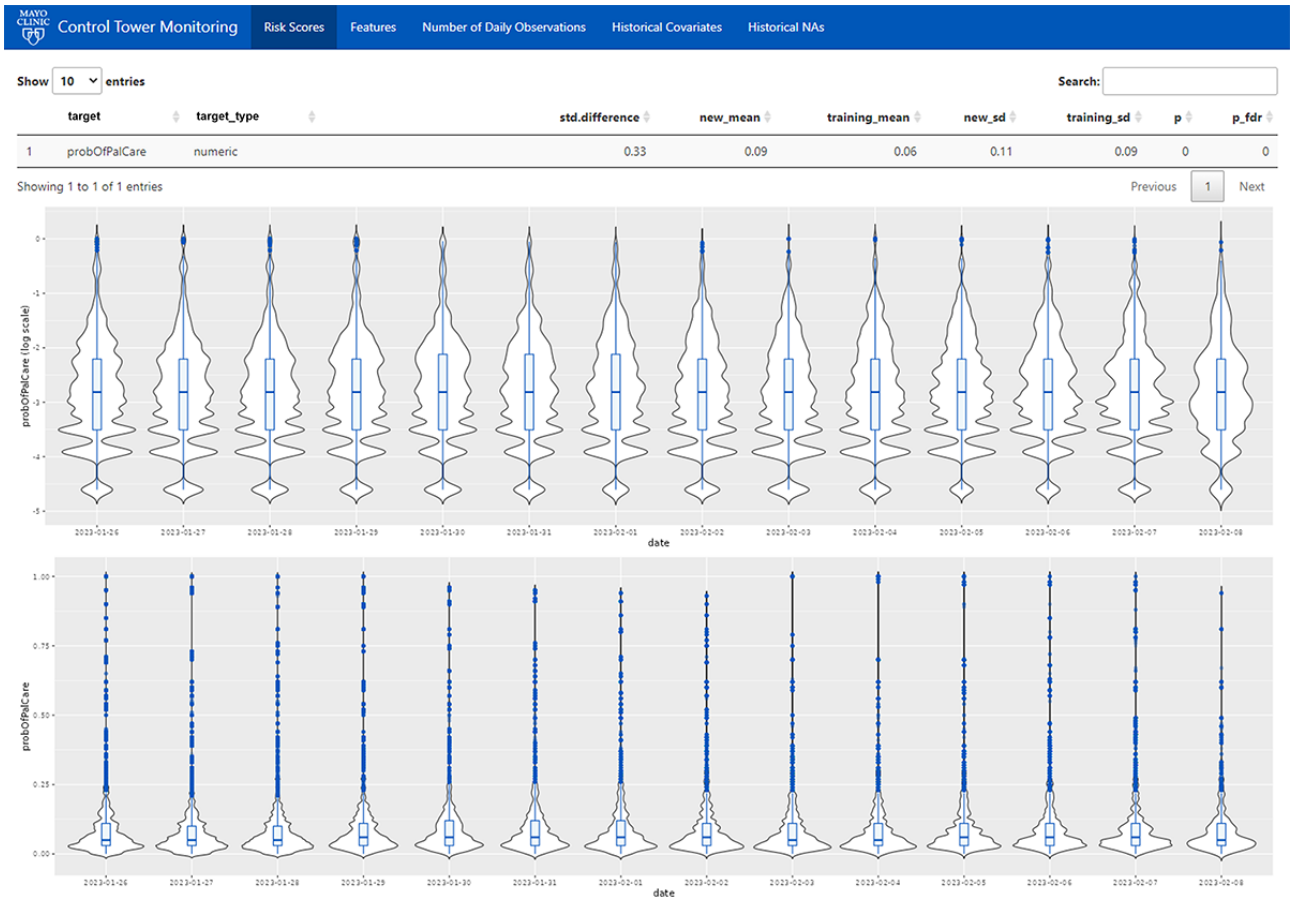
Figure 1. System architecture for Control Tower. For the Control Tower monitoring platform, we have 3 parent processes (training, production, and monitoring) that constitute our deployment. Child processes include the orchestration of the streams, events, and the prediction pipeline, which sends scores to the graphical user interface (GUI). EHR: electronic health record; REST API: Representational State Transfer application programming interface.



Monitoring Model Probabilities

In the absence of ground truth labels, predicted probabilities from the model were monitored as an alternative to evaluating model performance. The platform visualizes these predicted probabilities as distributions of daily risk scores (Figure 2). Distributions are plotted on probability and logarithmic scales, allowing for easier detection of shifts when most predicted probabilities are low, considering that most patients will not be "high risk." Historical daily distributions extend back 2 weeks, which is considered an optimal amount of time to notice shifts without overwhelming the user with data. Alongside these visualizations, several statistics are presented for comparing the prior 2 weeks data against the original training data. Means and SDs for the incoming and training distributions, the standard difference, and a *P* value for the Kolmogorov-Smirnov test of differences between the 2 distributions are provided. These statistics allow the user to detect gradual, more long-term changes that may go unnoticed when surveying 2 weeks of historic data. Overall, the tab shown in Figure 2 provides an overview of model predictions, allowing the user to quickly gauge whether a sudden or gradual probability shift has occurred.

Figure 2. The startup screen of the Control Tower performance monitoring platform. This screen provides the user a quick overview of the model's predicted probabilities over the past 2 weeks. The table near the top provides several statistics comparing the distribution of predicted probabilities over the last 2 weeks with the predicted probabilities on the training data. The 2 graphs contain a series of violin plots featuring the daily distribution of predicted probabilities. Given that the predicted probabilities cluster near 0, the distributions are also displayed on the log scale for easier visual inspection.



Monitoring Covariate Shift

Covariate shift was addressed in Control Tower by creating an interactive table containing all features included in the model (Figure 3) [4]. The table lists feature names and type, that is, continuous or discrete, and displays different statistical tests and plots, dependent on the feature type. To assess the impact of a feature with drift, the team included global feature importance scores from the originally trained model, in this case, the gradient boosting machine's relative influence rank statistic. Providing a ranking of features based on the extent of error reduction in the model enables the user to triage different drifts. All other things being equal, a drifting feature with higher importance to the model than another feature would indicate a higher priority need of a fix. Similar to the predicted probability tab, the previous 2 weeks of incoming data are compared with the training data, with standard differences, means, and SDs provided. To accommodate for the discrete variables present, the distributional Kolmogorov-Smirnov test is changed to the chi-square test. The user can sort the table by column, allowing them to quickly pinpoint features, for example, with high standard difference. Clicking on a feature's row in the table generates 2 plots underneath the table: the first is a line graph visualizing the daily standard differences, spanning back 2 weeks, and the second plot is dependent on the feature type. For continuous variables, the plot compares the feature's daily distributions over the past 2 weeks with the distribution of the training data, using box plots. For discrete variables, bar plots are displayed in a similar fashion indicating the percentage of patients where the discrete feature was present or "True." In tandem with the interactive table, these plots provide an efficient means of investigating a feature's historic values at a glance.

When a deeper investigation into a feature is necessary, the 2-week "look-back" may be insufficient. Therefore, the platform also keeps a log of the full historic feature trends, spanning back to when the model was deployed (Figure 4). Feature plots are sorted by the model's global feature importance and color-coded "green" or "red" to indicate whether the feature significantly drifted from the initial training distribution. Significance was determined via a nonparametric test developed by Capizzi and Masarotto [15], using a *P* value of .05. A nonparametric model was used because a moderate number of features were highly skewed, making traditional methods that assume normal distributions unworkable. Each feature contains plots for the location (level) and dispersion (scale) of the distribution. Overall, this tab, in addition to serving as a historical reference, provides a simple way to spot check for gradual shifts. Finally, an additional tab (Figure 5) is provided to assess the proportion of missing values over time, using the same visualizations and tests.

The final tab of the platform provides a simple line graph displaying the number of daily calls made to the model within the previous 2 weeks (Figure 6). Monitoring the number of daily

calls can provide quick insight into whether the model is performing appropriately. For example, an abnormal number of model calls in a day, such as 0, may indicate an error in the data pipeline or model environment.

Figure 3. The "Features" screen of the platform details the distributions of all features used by the model. The distribution of each feature based on the last 2 weeks of data is compared with the feature's distribution from the training data. These comparisons are provided via statistics in the table near the top, which can be sorted by each statistic to quickly find features with potential drift. Clicking on a feature populates 2 graphs, which are displayed below the table. The first graph displays the standardized difference between the feature's distribution for that day against the distribution from the training data. Below this graph, one of the 2 graphs will be displayed depending on whether the selected feature was binary or continuous. These graphs display the daily distributions of the feature, using bar graphs for binary features (red outline) or box plots for continuous features (yellow outline).

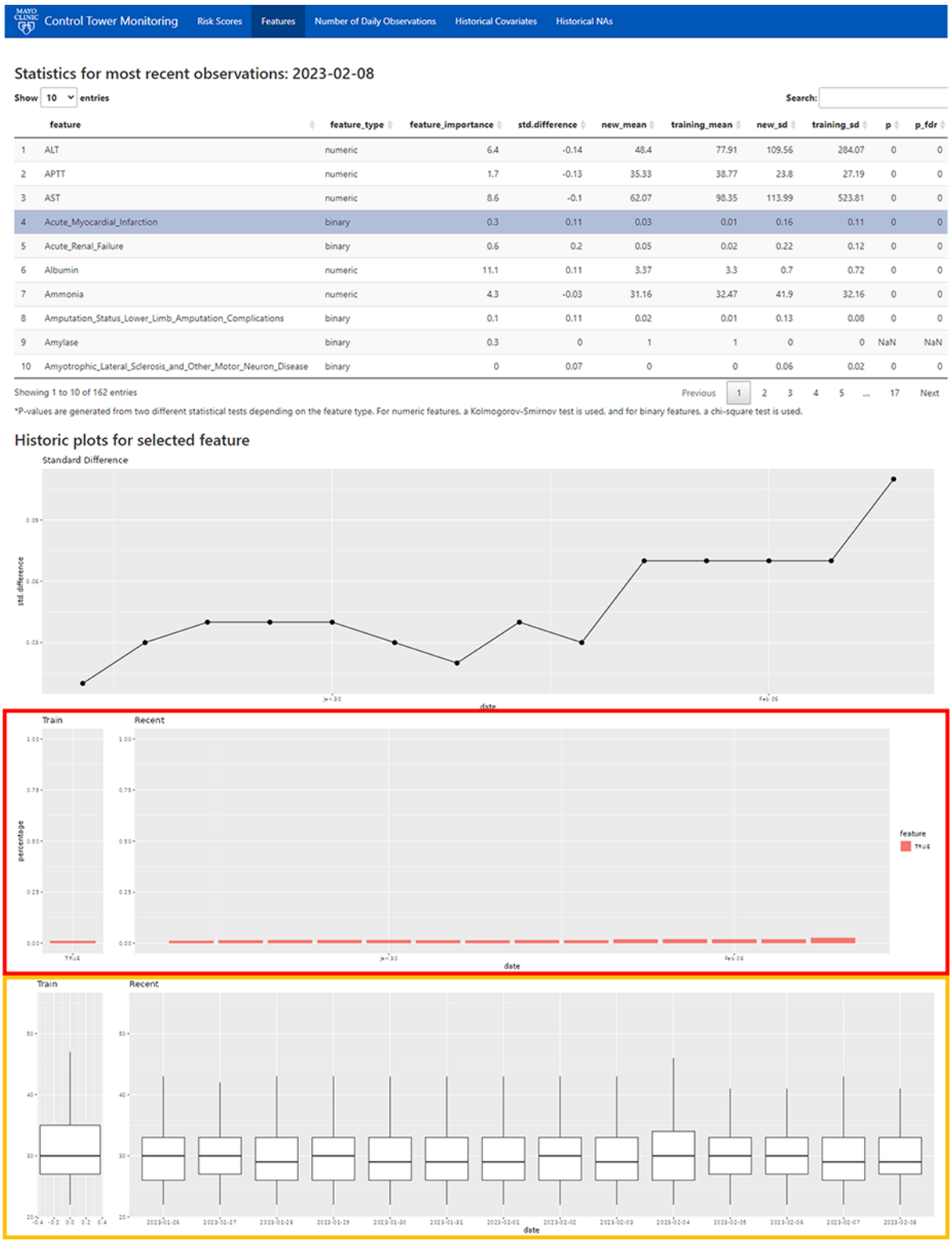


Figure 4. The "Historical Covariates" screen of the platform visualizes each feature's daily distribution, beginning with the training data and then spanning from the day the model was deployed and onward. Each feature contains plots for the location (level) and dispersion (scale) of the nonparametric distribution. Each feature's graph is color-coded "green" or "red" to indicate whether the feature's distribution has significantly drifted from the initial training distribution, with red indicating significant drift.

Figure 5. The "Historical NA's" screen of the platform visualizes each feature's historical missingness, beginning with the training data and then spanning from the day the model was deployed and onward. Each feature contains plots for the location (level) and dispersion (scale) of the nonparametric distribution. Each feature's graph is color-coded "green" or "red" to indicate whether the feature's missingness has significantly drifted from the initial training distribution, with red indicating significant drift. NA: not available.

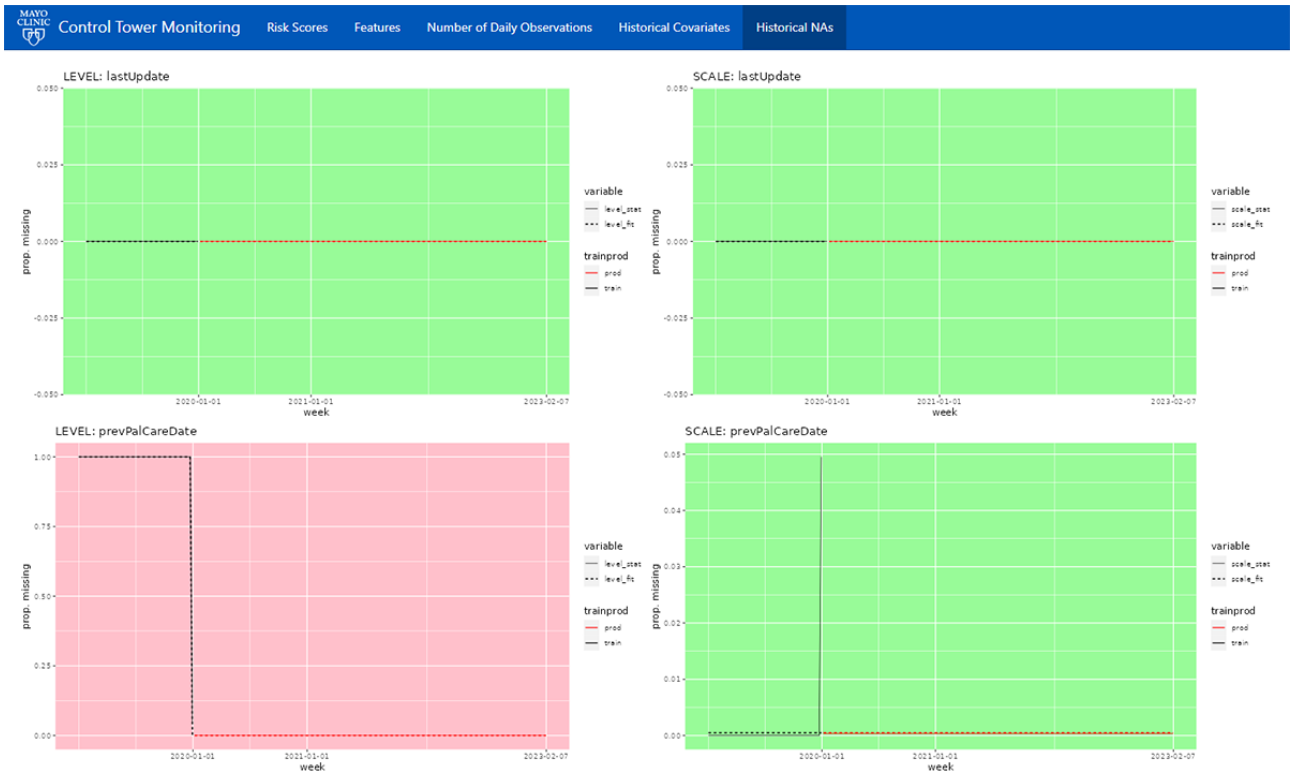
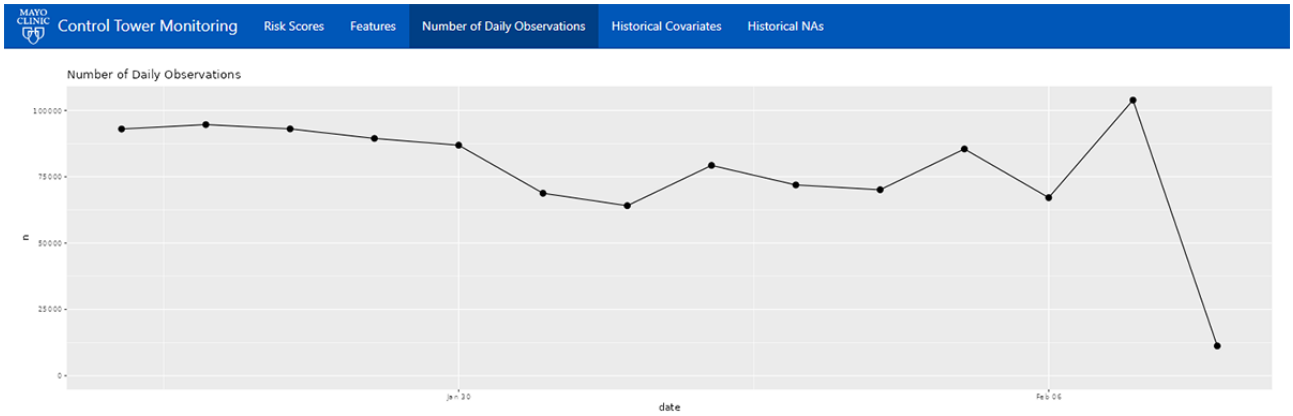


Figure 6. The “Number of Daily Observations” screen visualizes the number of model calls or predictions made each day over the past 2 weeks.



Error Classification

Any production-level model is susceptible to various errors, and Control Tower was no exception. Most errors primarily revolved around technical infrastructure, particularly issues with databases being inaccessible due to nightly processing or a high influx of requests. In Textbox 1, a sample of encountered errors while monitoring Control Tower is presented. Although some errors were seemingly random occurrences, such as server reboots or expired certificates, others were more frequent and persistent. For instance, every night at specific hours, the database that supplies data to Control Tower, called Clarity, became unresponsive due to data updates. On January 5, 2022, this process was delayed and caused errors in the morning scores. In addition, updates to our electronic health record (Epic, Epic Systems Corporation), often resulted in Clarity being temporarily unavailable. In such cases, most issues were resolved on the same day, requiring no further action besides acknowledging the possibility of outdated or missing scores. However, a few errors necessitated intervention. On November 7, 2022, a data mart containing diagnosis codes underwent structural changes, breaking a Control Tower query. Furthermore, the team identified a covariate shift where they observed a gradual decrease in troponin blood tests. This error was traced back to a change in laboratory codes used for troponin; the clinical practice had adopted a new laboratory code that was not present in the training data. To address this, the error was rectified by associating the new codes with the “Troponin” feature on the platform’s back end.

Textbox 1. Error logging: A convenience sample of encountered errors while monitoring Control Tower is presented. This was constructed through email chains of discussions between IT personnel who oversaw the Control Tower system and the data scientists who oversaw model delivery.

Date and error

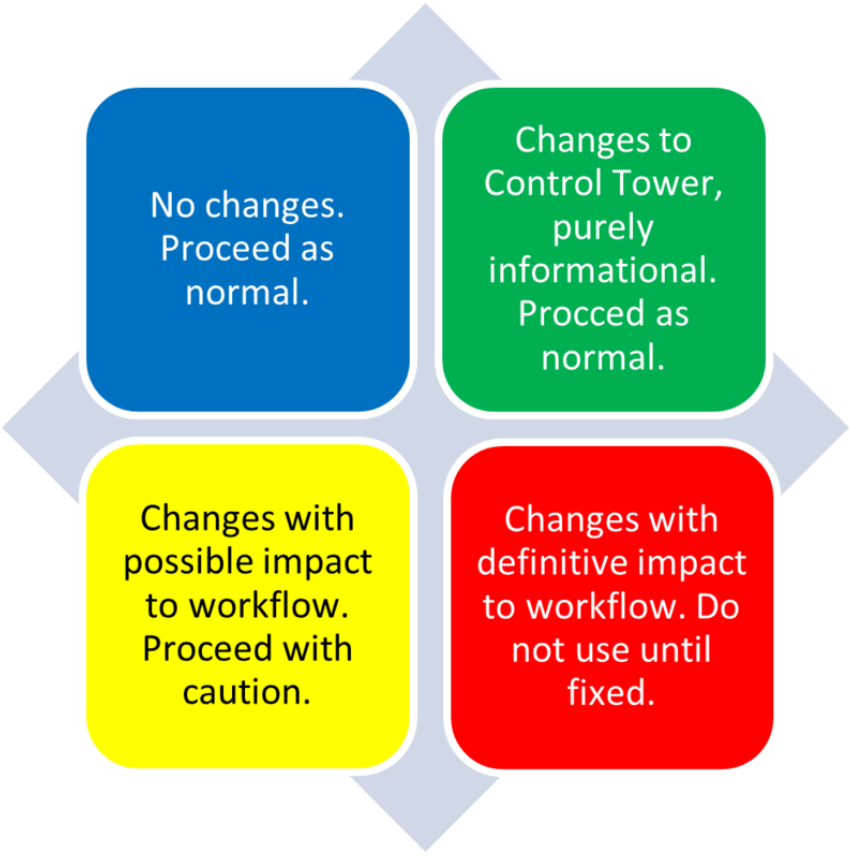
- August 26, 2019
  - Multiple errors in logs. It looks like calls were during 1:45 AM to 6:00 AM. During the period 1:45 AM to 6:00 AM, all messages are failing due to Clarity Refresh.
- November 27, 2019
  - Server reboot schedule, Control Tower team was not notified of schedule leading to unexpected downtime.
- July 24, 2020
  - Increase in FHIR (Fast Healthcare Interoperability Resources) API (application programming interface) for real-time observation calls leading to timeouts of model predictions.
- February 15, 2021
  - Generic FHIR API error call: “HTTP error code: 500.”
- April 8, 2021
  - Model errors after Epic upgrade.
- July 30, 2021
  - Troponin issues fixed causing covariate drift in model scores.
- September 15, 2021
  - Production system competed for resources requiring scale back of Control Tower scores updates. Errors created and schedule has now been updated for processing.
- January 5, 2022
  - Nightly Clarity Database delay causing morning score errors.
- May 16, 2022
  - IBM queue server certificate update causing server errors.
- November 7, 2022
  - Data mart for diagnoses codes update causing pipeline to break down.
- November 8, 2022
  - Issues with Clarity Database slowing down Control Tower queues.
- March 21, 2023
  - Control Tower FHIR API for real-time unit changes failing for a single request, causing payload slowdown.
- April 5, 2023
  - JSON structure changing causing model error (unintended repo change).
- May 1, 2023
  - An unplanned issue impacting Enterprise API Services, who manages real-time data feeds, resulting in internal server error.

Monitoring Protocol

Actions prompted by model monitoring feedback were synthesized into a protocol to communicate model failures and downtimes with clinical staff. The Control Tower protocol used a triage system, consisting of 4 stages in which each stage prompts a message to the clinical team, as outlined in Figure 7. The first stage (blue) was reserved for when everything was operating as expected. Stage 2 (green) indicated a minor change in the layout of the tool, such as a user interface change. These first 2 stages delivered informational prompts to the user, notifying them of the tool’s status and requiring no action from the user. The third stage (yellow) indicated possible performance degradation, such as when patient scores from the model were

not up-to-date, that is, a day old. The day-old scores can still provide evidence for action, but the clinical team may need to be cautious, as updated scores can change the patient's risk. As such, users were notified of these issues and asked to use the tool at their discretion. The last stage (red) indicated a significant error within the tool, such as a covariate completely missing from the model's input. This stage would notify staff not to use the tool until a fix was implemented.

Figure 7. Communication triage protocol for Control Tower. The protocol's stages are color-coded to signify different statuses and recommendations: blue for normal operation, green for minor layout changes, yellow for potential performance issues, and red for significant errors.



Design Considerations

Overview

From our experience of developing and implementing the Control Tower monitoring platform, we have derived a series of broader considerations necessary for model monitoring to serve as generalized guidelines for future implementations. Central to these guidelines arose themes of feasibility, design, implementation, and policy. While existing frameworks have proven successful in managing long-term IT infrastructure projects in health care, ML models are experimental and inherently open systems, entailing costly development and maintenance. As such, they demand additional considerations due to their reliance not solely on technical data but also on statistical and clinical assessments to identify errors. Consequently, there is no unequivocal, predetermined signal that can be provided to an IT group lacking clinical or data science expertise to detect these errors. Identifying them often necessitates the accumulation of statistical data over time. While readily available and accessible data may be used to identify some errors, others may require weeks or months of new data collection to draw inferences. Furthermore, in traditional software operations, refinements can be implemented more quickly. Responding to user feedback can often be met with bug fixes or minor feature requests. However, implementing refinements to an ML model often requires a longer development cycle, as many changes will require a complete retraining of the model or the acquisition of novel data. Such complications in model maintenance underscore the need for input from multiple teams, alongside established structures and policies, to ensure effective orchestration of model maintenance.

Feasibility: Are the Resources to Facilitate Productive Monitoring Available?

Successful real-world implementation of models must consider the present and *future* bandwidth limitations of a team. In an ideal scenario, a team would be provided ongoing dedicated time or personnel to ensure the upkeep of deployed models. However, this is not always feasible, and the responsibility of maintenance competes with a team's constant stream of new projects and tasks. As such, it is important to first determine how long-term monitoring of a model (or eventually, *models*) will be integrated into the team's workflow. For example, who will check in on the model and with what frequency? If and when the model requires retraining, who will perform the retraining, and how will they be guaranteed the flexibility to shift from their current projects to accomplish this?

In addition to personnel, computational resources must also be considered for long-term monitoring. Regarding storage requirements, the amount of data produced from monitoring

depends on a variety of factors, including the model's feature set, the frequency of calls made to the model, and the number of feature and performance metrics that will be tracked. For example, a model that delivers constant, real-time feedback in a critical care setting may, in turn, require constant monitoring to ensure performance does not suddenly degrade, resulting in performance metrics and feature distribution logs needing to be generated continuously.

When assessing the feasibility of long-term monitoring, an attractive option to consider is automated monitoring: developing models that detect whether a significant change has occurred in a deployed model's predictions or its incoming data. While our team chose a "hands-on" approach, others have found success in implementing an additional model for performance monitoring to notify the team of data shifts and, in some cases, automatically retrain the original model [16]. Ultimately, the decision to use an automated monitoring model comes down to the type of model deployed, the bandwidth of the team, and the reliability of the data to support automatically retraining the deployed model. In any case, even an automated monitoring model will still require some human monitoring as well. Our team is currently working on an automated monitoring system for this purpose.

Design: Deciding What and How to Monitor?

Overview

The design of an investigative platform to facilitate model monitoring may range from dynamic and interactive interfaces to static reports, the choice of which is dependent on feasibility factors and nuances of the particular scenario, but design should ultimately enable rapid and comprehensive assessment. Furthermore, there are standard functionalities each platform should feature to appropriately assess long-term model performance.

Deployed models encounter performance declines through distributional and relational shifts in the underlying data [17]. These shifts are the crux of why postdeployment monitoring is necessary, and no model is immune to them, regardless of how well it performed in its testing environment [18]. This impediment has received a wealth of attention from the ML community and has been synthesized into 3 types of distinct shifts.

Prior Probability Shift

The distribution of the target variable Y changes between the training data and incoming data, but not P (X|Y) [19]. This can occur when the prevalence of a disease changes over time in the target population; however, the underlying factors that cause the disease remain constant, for example, a spike in influenza rates during the influenza season. Prior probability shift is assessed by monitoring the distribution of the target variable over time, measuring for any sudden or gradual changes.

Covariate Shift

Distributions of input data diverge between training data and new incoming data [4]. Such shifts may occur in the clinical setting, for example, when diagnostic screenings are updated. This procedural change may decrease the specific laboratory values, which are heavily relied upon by the model. Conversely, diagnostic variables that were initially infrequent may become more prevalent over time. This can result in situations where the model, which had limited instances of these variables during training, struggles to fully capture, and therefore use, their predictive signals.

Concept Drift

The relationship between the incoming input data and target variable changes over time, drifting from the original relationship captured in the training data [20]. The COVID-19 pandemic provided a real-world example of concept drift, as hospital census models were affected by admissions that drastically moved toward higher-risk patients due to increases in complications from the COVID-19 disease and decreases in hospital use among people with a milder spectrum of illness [21].

Usability

A successful UI will take into consideration the professional backgrounds of those using the platform. However, when the responsibilities of monitoring are handed to a different group, the new group's level of familiarity with the model should guide the design. For example, guidelines for what is acceptable variance should be established and implemented. One method for accomplishing this may be through using control charts, allowing the modeling team to prespecify a simple and visual approximation of how much drift is tolerable before action must be taken [22].

Implementation: How Will the Platform Be Built and Sustained?

When implementing a monitoring platform, it is necessary to consider how the back end of the platform will process and store the necessary data elements. The efficiency of this task is critical and must accommodate the model's scale and responsiveness. Data can amass quickly as large feature sets are monitored, and the model may be called frequently to predict on many patients throughout the day. Furthermore, the back end must be capable of efficiently parsing, formatting, and, if necessary, compressing the data into clean data sets for the platform to analyze and visualize. For Control Tower, many of these data storage requirements were already in place for capturing and storing the necessary patient elements. This will likely be the case for many clinical scenarios, as patient data must be securely and efficiently housed. Instead, implementation efforts are more apt to center around ensuring these data elements are efficiently piped to the monitoring platform.

Using a web application for model monitoring provides a dynamic interface, allowing any user with log-in permissions to view the real-time status of the model and the surrounding data. This investigation mechanism can eliminate potential confusion, which may arise from a routine generation and sharing of static technical reports, such as accidentally referencing outdated documents. When selecting a programming language to build the app, preference should be given to those languages that facilitate efficient app development. For Control Tower, R Shiny was used given the team's previous experience with the package and strong background in R. The R package

provides a user-friendly environment for quickly creating, testing, and publishing web applications. Similar web application tools exist across multiple programming languages including Python and Java, and as such, teams are likely to find a web application package in a language they are familiar with.

When coding the app, modular coding practices should be adopted to ensure flexibility and scalability. Such adoption promotes versatility of the app to incorporate additional statistical measures or visualizations and allows the app to be easily translated for other monitoring use cases. Leveraging modular coding practices at the onset of app development allows for future additions, revisions, and ports to be made with minimal effort. For Control Tower, modular coding practices were primarily used to better facilitate development across multiple team members. This practice allowed for functions to be easily repurposed by other team members to avoid duplication of work and to allow the app to be easily extended to other ML models within Mayo Clinic.

The number of programming languages used in the data pipeline plays a significant role in shaping the development process and the overall efficiency of the monitoring. To facilitate this, minimizing the number of programming languages used across the various tasks can streamline development and maintenance through ease of interpretability and integration. This can reduce maintenance costs and overhead by reducing interoperability concerns and decreasing the learning curve for new team members. Minimizing these ongoing costs is a necessity when considering the model will ideally be in production long term. However, if the development team is proficient in multiple languages, leveraging the strengths of each may have its advantages, such as reducing bottlenecks in development or data transfer, while increasing the flexibility of a system. In the case of Control Tower, R (R Foundation for Statistical Computing), Python (Python Software Foundation), and shell scripts were used, favoring R for app development, Python for data processing, and shell scripts for scheduling various model and platform tasks.

In addition, upstream problems will inevitably manifest; therefore, implementing a notification system for these errors can proactively address disruptions, minimizing the downtime of the pipeline. One method for accomplishing this is to incorporate error logging and alerts into cron jobs, which can immediately notify the team of any failures. Such notifications are critical for model monitoring, as some errors may be undetectable to the end user, resulting in the continued use of inaccurate information. As such, it is vital for monitoring teams to identify, communicate, and resolve errors as soon as possible.

Finally, integrating regular checking of the platform into the team workflow allowed the team to not only stay abreast of model performance but also maintain an intuitive sense of potential broader complications surrounding the model. For example, monitoring the probability distributions of the model ultimately provided the team with a sense of whether further investigations into the model would be necessary. However, investigations into the distributions of the individual features allowed for potential diagnoses as to why the model may begin to degrade in performance, as well as alluded to data pipeline errors that may be present. By maintaining a sense of these wider issues, shifts in the outcome could be more easily prevented and diagnosed

Policy: What Is the Response to Platform Feedback?

Overview

Once the monitoring platform is deployed and available, the next stage of considerations surrounds how knowledge provided by the platform will be used. A set of policies must be developed to determine which actions will be taken based on monitoring feedback, addressing such questions as “At what point is a data shift significant enough to prompt retraining?” and “How will errors be communicated with technical and clinical staff?” Generally, such a policy should cover error designation and response, when to retrain, and how to communicate failures. In addition, a well-defined policy allows for the task of monitoring to more easily be extended across various teams and roles.

Error Designation and Response

It is essential to establish and define a process that determines when a specific degree of shift or drift in the model qualifies as an error warranting a response. The question *“How much drift is necessary to take action?”* represents one of the more subjective aspects of model monitoring. In scenarios where multiple team members are tasked with overseeing model performance or possess limited familiarity with the model, substantial interrater variability becomes a concern. For example, one team member might observe a 5% shift in the distribution of a feature and consider it inconsequential, while another member might view it as a reason for immediate action. To address this variability, the Control Tower team would send email updates to other team members detailing any shifts that were noticed; this would allow for a collective discussion on whether to take action as well as allow for a convenient forum to keep all team members updated on the model’s status. Regardless of the criteria used to identify shifted covariates or outcomes, team members must communicate and establish agreement on the minimal drift threshold requiring action, while ensuring that utmost priority is placed on maintaining optimal model accuracy.

Even with consensus on the magnitude of a shift, several contextual factors can influence the team’s risk tolerance toward these shifts. Significant changes may occur without sustained trends, indicating a regression to the mean. Alternatively, a dramatic shift might happen for a variable with minimal contribution to the risk score. While predefined cutoff points could be considered to standardize investigations, these benchmarks may still necessitate ongoing human review and could vary for each feature, making it impractical to define for every feature in large feature sets.

Even if an error is defined with a certain level of risk in mind, there are considerations in the response to the error and the amount of time one needs to allocate for remediation. A deployed model is prone to errors from a variety of sources, ranging from data shifts to IT scheme modifications. Given the diversity of potential errors, an effective policy will include guidelines for the categorization of errors along with the

appropriate responses to each. The errors encountered with Control Tower fell broadly into 4 categories.

1. Technical infrastructure: database issues, expiration of certificates, and password updates often causing the pipeline to fail
2. Explained shift: a significant data shift with an identified root cause
3. Unexplained shift: a significant data shift with an unidentified root cause
4. Performance loss: a decrease in the model's performance metrics, which may manifest with or without data shift

Categorizing errors for appropriate response is crucial, as it establishes a standardized knowledge base for reporting, ultimately enhancing the efficiency of troubleshooting. Categorization often leads to the discovery of similar strategies for mitigating similar error types. For instance, errors related to database refreshing or password expiration typically do not require immediate intervention, while performance losses in accuracy or calibration often necessitate retraining of the model. Appropriate categorization also offers the advantage of reducing risk tolerance while enhancing response efficiency. Having encountered an error previously increases the likelihood of streamlining investigations, enabling the examination of lower-risk shifts or drifts.

When ongoing outcomes data are available, performance loss can be detected by looking for significant shifts across a variety of classification performance metrics including area under the receiver operating characteristic, area under the precision-recall curve, calibration, subgroup differences, and so on. When such data are absent, as in the case of Control Tower, performance loss can only be inferred by looking for significant shifts in the distribution of predicted probabilities of the model. To supplement assessing predicted probabilities, potential performance loss may also be identified by looking for significant shifts in the features of the model. While significant shifts may occur in these features without significant shifts in the model's output, drifts in feature distributions can signal other potential problems necessary to address. While performance loss may be resolved or mitigated through upstream pipeline errors, some instances may require the model to be retrained.

Model Retraining

The circumstances for when to retrain a model will vary across teams and platforms, often dependent on the cost of retraining. As such, it is necessary for a platform policy to clearly state when, and when not, to retrain. For example, many errors will not require model retraining such as simple pipeline errors or data shifts due to changes in medical coding, requiring only a small update to the pipeline. Therefore, it is important to first identify and fix any upstream errors before considering retraining. There are even instances of significant shifts that do not warrant retraining. For example, one could have several shifted covariates in the model with trivial importance scores, effectively having no impact on predictive performance. From the perspective of model importance, one may bin covariates that have little impact and essentially treat them as nuisance variables.

Assuming no upstream errors are present, a model should always be retrained when significant and sustained performance loss is encountered. Defining *significant* and *sustained* will be specific to each scenario, depending on the algorithm and health care delivery model. However, it is incumbent upon the team to define an appropriate window for performance to vary, with a lower limit triggering retraining.

It is important to note that retraining does not have to be used sparingly, assuming the bandwidth is available. When feasible, it may be good practice to routinely retrain the model with the expectation that updated data are more current with clinical practice. Such versioning of the model would allow for new features, incremental improvements, and technical debt management. For Control Tower, versioning allowed us to spot potential bugs or fixes and investigate new features.

Communicating With Clinical Staff

The clinical team using the model's outputs must be consistently informed about the model's status due to its significance to their workflow and overall trust in the model. The model's standard operating procedure outlines how the clinical team should use the model and details communication protocols between IT, data science, and the clinical users. Protocols should consist of dedicated contacts for various issues and plans for how to operate during model performance shifts and downtime.

Discussion

Principal Findings

As ML models require consistent monitoring to ensure sustained accuracy, a series of decisions must be made for how best to integrate model monitoring into a team's workflow. Problems, considerations, and solutions that arise from this process can vary greatly depending on the setting, nature of the model, and available bandwidth, both from the technical team and their computational resources. While prior work has established the importance of monitoring and corresponding statistical solutions, this paper provides specific considerations and solutions derived from the real-world implementation and day-to-day use [23,24]. Throughout the integration of Control Tower, our team found that these considerations centered around 4 phases that serve as a road map when planning a long-term modeling strategy: feasibility, design, implementation, and policy.

Experiences

Development and implementation of the platform faced several obstacles, which we attribute to the inherent realities of integrating real-world applications. First, the team was unable to complete the platform by the time the associated clinical trial for the Control Tower model began recruitment. This required the team to omit crucial features from the platform, such as monitoring for concept drift. Monitoring concept drift required collecting ground truth outcomes, that is, whether patients actually received palliative care. Collecting these patient outcomes required building a separate data pipeline, which was the team's original intent, but as the team took on additional tasks, the pipeline was passed over in favor of monitoring predicted probabilities. While omitted from Control Tower in scenarios where outcomes data are tracked, we direct the readers

to literature providing a more comprehensive understanding of concept drift [25-27].

The original intention for Control Tower was to have the model run any time a patient's laboratory values were updated, ensuring that the patient's risk scores were always reflective of current data. While the model was originally deployed using this dynamic system, it, unfortunately, proved too taxing for the IT infrastructure in which it was hosted. To alleviate this problem, the model and platform were switched to running on a batch schedule, updating patient risk scores and the monitoring platform every 4 hours. While this delivery schedule proved more manageable, model calls made between these 4-hour updates ran the risk of using outdated patient data, potentially impacting performance. Given that the workload imposed by the original schedule was infeasible, this was considered a fair compromise. Finally, the implementation of the platform occurred during the COVID-19 pandemic, which affected staffing and resulted in IT furloughs. Unfortunately, this meant that technical infrastructure problems, which could typically be fixed by IT on the same day, instead, took up to 1 week to fix, resulting in prolonged downtime for the model.

Despite these challenges, there were several positive experiences to highlight. First, a significant amount of collaboration occurred within the data science team in order to have the monitoring platform in a usable state by the time of the model's clinical trial. This required analysts to tend to a variety of tasks, often on a moment's notice. Following deployment, there was also sufficient bandwidth from the team members to continue monitoring the platform as they took on additional projects. Second, IT furloughs as a result of the pandemic were resolved within 6 months, allowing routine technical infrastructure issues to once again be resolved on the day of occurrence, resulting in less model downtime. Finally, the model's predicted probabilities remained, for the most part, consistent, making for a stable tool throughout the documented 2 years of use. Using a simple linear regression model, we examined the relationship between daily predicted probabilities (dependent variable) and time since deployment (independent variable), observing a slope of 0.005 at a P value of <.001, suggesting a statistically significant, but functionally small trend, with the mean probability increasing .005% each day.

Limitations

Despite a thorough detailing of our experiences, it is important to note that this paper covers only a single implementation. While we have recounted the challenges and considerations necessary for Control Tower, this is not an exhaustive list, and other teams and platforms may encounter challenges foreign to ours.

Future Considerations

The Control Tower platform allowed our team to successfully monitor performance and maintain our deployed model for 2 years. Moving forward, our team is planning to automate parts of this task, for example, by implementing an automated email notification system, notifying the team when the number of model calls, predicted probabilities, and incoming data streams shift beyond their respective significance thresholds. While this modification is not intended to outright replace the manual checks of the platform, it will allow the team to check the platform at a lesser frequency. This system will serve as a placeholder while the team develops a new model to monitoring the performance of Control Tower, leveraging supervised learning to detect shifts in the probability and multivariate covariate distributions [28,29].

The team also considered an online or continuous learning model to automatically address data drift. In continuous learning, the algorithm would update its predictions as new data come in and alleviates the need to manually retrain the data [30]. Although appealing, an automated system, in this sense, would require more policy changes and would bring with it a number of issues. First, there are several cost and computing issues that could make an implementation difficult, as entire systems would need to know when to train and to do it without interrupting the current pipeline, as well as a validation step to ensure sustained, if not improved, accuracy. Second, the algorithm must remain trustworthy for clinicians. Did the algorithm *unlearn* anything important? Did it learn anything irrelevant or incorrect? As an example, if a covariate shift occurred due to a missing laboratory code, resulting in increasingly missing values of that laboratory, we would not want the model to learn a new relationship with the missingness; instead, we would make an update to the data pipeline to resolve the missingness. Finally, all continuous learning models require ready access to the gold-standard outcome, which might not be feasible in all cases.

Conclusions

Once an ML model has been successfully developed and deployed, it must be continuously monitored to ensure its efficacy amidst an ever-evolving practice and stream of patients. While a variety of methods have been proposed to statistically monitor the performance of models, this is only one factor to consider when implementing a long-term modeling strategy. By disseminating the broader experiences of integrating ML monitoring platforms into clinical practice, readers will be better equipped for the considerations and challenges encountered during their own implementations.

Acknowledgments

The research reported in this paper was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (award number R01EB019403). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to concerns of patient identification but are available from the corresponding author on reasonable request. Such requests may require separate approval by the Mayo Clinic Institutional Review Board.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Source code for a demo of the Control Tower Model Monitoring R Shiny application.
[ZIP File (Zip Archive), 26937 KB - medinform_v12i1e50437_app1.zip]

References

1. Allen B, Dreyer K, Stibolt RJ, Agarwal S, Coombs L, Treml C, et al. Evaluation and real-world performance monitoring of artificial intelligence models in clinical practice: try it, buy it, check it. J Am Coll Radiol 2021 Dec;18(11):1489-1496. [doi: 10.1016/j.jacr.2021.08.022] [Medline: 34599876]
2. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med 2021 Aug 01;181(8):1065-1070 [FREE Full text] [doi: 10.1001/jamainternmed.2021.2626] [Medline: 34152373]
3. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med 2021 Jul 15;385(3):283-286 [FREE Full text] [doi: 10.1056/NEJMc2104626] [Medline: 34260843]
4. Schwaighofer A, Quinonero-Candela J, Sugiyama M, Lawrence ND. Dataset Shift in Machine Learning. New York, NY: Penguin Random House LLC; 2008.
5. Klinkenberg R, Joachims T. Detecting concept drift with support vector machines. In: Proceedings of the Seventeenth International Conference on Machine Learning. 2000 Presented at: ICML 2000; June 29-July 02, 2000; Stanford, CA. [doi: 10.1007/978-1-4615-0907-3_3]
6. Huang R, Geng A, Li Y. On the importance of gradients for detecting distributional shifts in the wild. arXiv Preprint posted online October 1, 2021.
7. Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization. 2021 Jun 28. URL: https://www.who.int/publications/i/item/9789240029200 [accessed 2023-07-05]
8. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. In: Proceedings of the Software Engineering for Machine Learning. 2014 Presented at: SE4ML 2014; December 13, 2014; Montreal, QC. [doi: 10.7551/mitpress/12440.003.0011]
9. Pruneski JA, Williams RJ3, Nwachukwu BU, Ramkumar PN, Kiapour AM, Martin RK, et al. The development and deployment of machine learning models. Knee Surg Sports Traumatol Arthrosc 2022 Dec;30(12):3917-3923. [doi: 10.1007/s00167-022-07155-4] [Medline: 36083354]
10. Murphree DH, Wilson PM, Asai SW, Quest DJ, Lin Y, Mukherjee P, et al. Improving the delivery of palliative care through predictive modeling and healthcare informatics. J Am Med Inform Assoc 2021 Jul 12;28(6):1065-1073 [FREE Full text] [doi: 10.1093/jamia/ocaa211] [Medline: 33611523]
11. Wilson PM, Philpot LM, Ramar P, Storlie CB, Strand J, Morgan AA, et al. Improving time to palliative care review with predictive modeling in an inpatient adult population: study protocol for a stepped-wedge, pragmatic randomized controlled trial. Trials 2021 Oct 16;22(1):635 [FREE Full text] [doi: 10.1186/s13063-021-05546-5] [Medline: 34530871]
12. Wilson PM, Ramar P, Philpot LM, Soleimani J, Ebbert JO, Storlie CB, et al. Effect of an artificial intelligence decision support tool on palliative care referral in hospitalized patients: a randomized clinical trial. J Pain Symptom Manage 2023 Jul;66(1):24-32. [doi: 10.1016/j.jpainsymman.2023.02.317] [Medline: 36842541]
13. Dalkey NC, Brown BB, Cochran S. The Delphi Method: An Experimental Study of Group Opinion. Santa Monica, CA: RAND Corporation; 1969.
14. Barrett D, Heale R. What are Delphi studies? Evid Based Nurs 2020 Jul 19;23(3):68-69. [doi: 10.1136/ebnurs-2020-103303] [Medline: 32430290]
15. Capizzi G, Masarotto G. Phase I distribution-free analysis of univariate data. J Qual Technol 2017 Nov 21;45(3):273-284. [doi: 10.1080/00224065.2013.11917938]
16. Shayesteh B, Fu C, Ebrahimzadeh A, Glitho RH. Automated concept drift handling for fault prediction in edge clouds using reinforcement learning. IEEE Trans Netw Serv Manage 2022 Jun;19(2):1321-1335. [doi: 10.1109/tnsm.2022.3153279]
17. Nestor B, McDermott MB, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. arXiv Preprint posted online August 02, 2019 [FREE Full text]

18. Fu S, Wen A, Schaeferle GM, Wilson PM, Demuth G, Ruan X, et al. Assessment of data quality variability across two EHR systems through a case study of post-surgical complications. AMIA Jt Summits Transl Sci Proc 2022;2022:196-205 [FREE Full text] [Medline: 35854735]
19. Dockès J, Varoquaux G, Poline JB. Preventing dataset shift from breaking machine-learning biomarkers. Gigascience 2021 Oct 28;10(9):giab055 [FREE Full text] [doi: 10.1093/gigascience/giab055] [Medline: 34585237]
20. Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TW, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Sci Rep 2021 Dec 26;11(1):23017 [FREE Full text] [doi: 10.1038/s41598-021-02481-y] [Medline: 34837021]
21. Moynihan R, Sanders S, Michaleff ZA, Scott AM, Clark J, To EJ, et al. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. BMJ Open 2021 Mar 16;11(3):e045343 [FREE Full text] [doi: 10.1136/bmjopen-2020-045343] [Medline: 33727273]
22. Woodall WH, Spitzner DJ, Montgomery DC, Gupta S. Using control charts to monitor process and product quality profiles. J Qual Technol 2018 Feb 16;36(3):309-320. [doi: 10.1080/00224065.2004.11980276]
23. Petersen C, Smith J, Freimuth RR, Goodman KW, Jackson GP, Kannry J, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. J Am Med Inform Assoc 2021 Mar 18;28(4):677-684 [FREE Full text] [doi: 10.1093/jamia/ocaa319] [Medline: 33447854]
24. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based Software as a Medical Device (SaMD) - discussion paper and request for feedback. U.S. Food & Drug Administration. URL: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf [accessed 2023-07-05]
25. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: a review. IEEE Trans Knowl Data Eng 2018 Oct 18;31(12):2346-2363. [doi: 10.1109/tkde.2018.2876857]
26. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv 2014 Mar;46(4):1-37. [doi: 10.1145/2523813]
27. Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Japkowicz N, Stefanowski J, editors. Big Data Analysis: New Algorithms for a New Society. Cham, Switzerland: Springer; 2016.
28. Hwang W, Runger G, Tuv E. Multivariate statistical process control with artificial contrasts. IIE Transact 2007 Mar 22;39(6):659-669. [doi: 10.1080/07408170600899615]
29. Deng H, Runger G, Tuv E. System monitoring with real-time contrasts. J Qual Technol 2017 Nov 21;44(1):9-27. [doi: 10.1080/00224065.2012.11917878]
30. Lee CS, Lee AY. Clinical applications of continual learning machine learning. Lancet Digit Health 2020 Jul;2(6):e279-e281 [FREE Full text] [doi: 10.1016/S2589-7500(20)30102-3] [Medline: 33328120]

Abbreviations

GUI: graphical user interface
ML: machine learning

Edited by C Lovis; submitted 07.07.23; peer-reviewed by H Joo, C Yu, M Bjelogrlic, H Mueller; comments to author 03.08.23; revised version received 22.08.23; accepted 04.05.24; published 28.06.24.

Please cite as:
Faust L, Wilson P, Asai S, Fu S, Liu H, Ruan X, Storlie C
Considerations for Quality Control Monitoring of Machine Learning Models in Clinical Practice
JMIR Med Inform 2024;12:e50437
URL: https://medinform.jmir.org/2024/1/e50437
doi:10.2196/50437
PMID:

Implementing a Biomedical Data Warehouse From Blueprint to Bedside in a Regional French University Hospital Setting: Unveiling Processes, Overcoming Challenges, and Extracting Clinical Insight

Matilde Karakachoff[1], MSc; Thomas Goronflot[1], MSc; Sandrine Coudol[1], MSc; Delphine Toublant[1,2], MSc; Adrien Bazoge[1,3], MSc, PhD; Pacôme Constant Dit Beaufils[1,4], MD; Emilie Varey[1,5], MSc; Christophe Leux[6], MD, PhD; Nicolas Mauduit[6], MD; Matthieu Wargny[1], MD, PhD; Pierre-Antoine Gourraud[1,7], MPH, PhD

1

2

3

4

5

6

7

Corresponding Author:
Pierre-Antoine Gourraud, MPH, PhD

Abstract

Background: Biomedical data warehouses (BDWs) have become an essential tool to facilitate the reuse of health data for both research and decisional applications. Beyond technical issues, the implementation of BDWs requires strong institutional data governance and operational knowledge of the European and national legal framework for the management of research data access and use.

Objective: In this paper, we describe the compound process of implementation and the contents of a regional university hospital BDW.

Methods: We present the actions and challenges regarding organizational changes, technical architecture, and shared governance that took place to develop the Nantes BDW. We describe the process to access clinical contents, give details about patient data protection, and use examples to illustrate merging clinical insights.

Implementation (Results): More than 68 million textual documents and 543 million pieces of coded information concerning approximately 1.5 million patients admitted to CHUN between 2002 and 2022 can be queried and transformed to be made available to investigators. Since its creation in 2018, 269 projects have benefited from the Nantes BDW. Access to data is organized according to data use and regulatory requirements.

Conclusions: Data use is entirely determined by the scientific question posed. It is the vector of legitimacy of data access for secondary use. Enabling access to a BDW is a game changer for research and all operational situations in need of data. Finally, data governance must prevail over technical issues in institution data strategy vis-à-vis care professionals and patients alike.

(JMIR Med Inform 2024;12:e50194) doi:10.2196/50194

KEYWORDS

data warehouse; biomedical data warehouse; clinical data repository; electronic health records; data reuse; secondary use; clinical routine data; real-world data; implementation report

Introduction

The increasing use of electronic health records in research settings presents physicians with the systematic yet secondary use of data collected from multiple sources [1-3]. Indeed, hospital information systems (HISs) face a technical challenge to harmonize and integrate application systems and clinical databases that are highly heterogeneous, are based on editor-specific software formats, and use nonstandardized terminologies [3,4].

Moreover, institutions must face the legal and ethical challenge of granting secondary access to data due to national and international laws vis-à-vis patient privacy [5-7]. The reuse of data produced during the care process implies operational knowledge of ethics and legacy concepts that must be solved through well-defined data governance and access policies [7].

Repositories must ensure not only the technical aspects to data access but also the decision criteria granting access [6]. Indeed, institutional data governance is also pivotal when considering legal and ethical principles such as patient informed consent and privacy data protection [2,8,9]. Last but not least, following the line of traditional epidemiology, clinical data reuse requires treatment within a validated and standardized methodological framework to ensure a qualitative result from a scientific and clinical point of view [9].

Despite these difficulties, the rise of biomedical data warehouses (BDWs) is transforming research processes for epidemiology and clinical studies [5,10-12]. Patient data constitute well-defined profiles that can be used to facilitate the enrichment of cohorts [13,14], patient selection and follow-up for clinical trials [15], phenotypic detection, and detailed descriptions of symptoms. BDWs can facilitate the development and performance of personalized and precision medicine, including through the use of big data and artificial intelligence methods.

The implementation of BDWs is conducted at various geographical levels in France. To our knowledge, 24 active hospital BDWs were set up between 2008 and 2023 [16-21]. Regional coordination often occurs within specialized networks of BDWs such as the "Ouest Data Hub," which is specifically designed for university hospitals in Western France [22]. This can take place in thematic networks and is well advanced in cancer data [23]. National or European Union–wide initiatives also propose a development and coordination framework to deal with the different challenges of implementation. In particular, France initiated in 2019 a national project called "Health Data Hub" [24,25] that promotes centralized coordination and increases the visibility of data sources on a nationwide level.

In this paper we report our 5-year experience regarding organizational changes, technical architecture, and governance, supporting the implementation of the Nantes University Hospital Biomedical Data Warehouse (NBDW). We describe the process to access clinical content. We also give figures concerning sources and contents included in the repository, and provide some insight into the projects to which the NBDW contributed. Finally, we propose three indicators to measure the effectiveness of the setting up operation.

Methods

Overview

In 2018, Centre Hospitalo-Universitaire de Nantes (CHUN; University Hospital of Nantes) implemented a BDW to facilitate secondary use of personal health data originally collected in the context of patient care for research and to offer single secure access to up-to-date data from different sources within the CHUN HIS, accommodating a wide range of data types, including demographic and clinical information, consultations, billing codes, diagnoses, laboratory results, medical notes, and drug administration in a unified view.

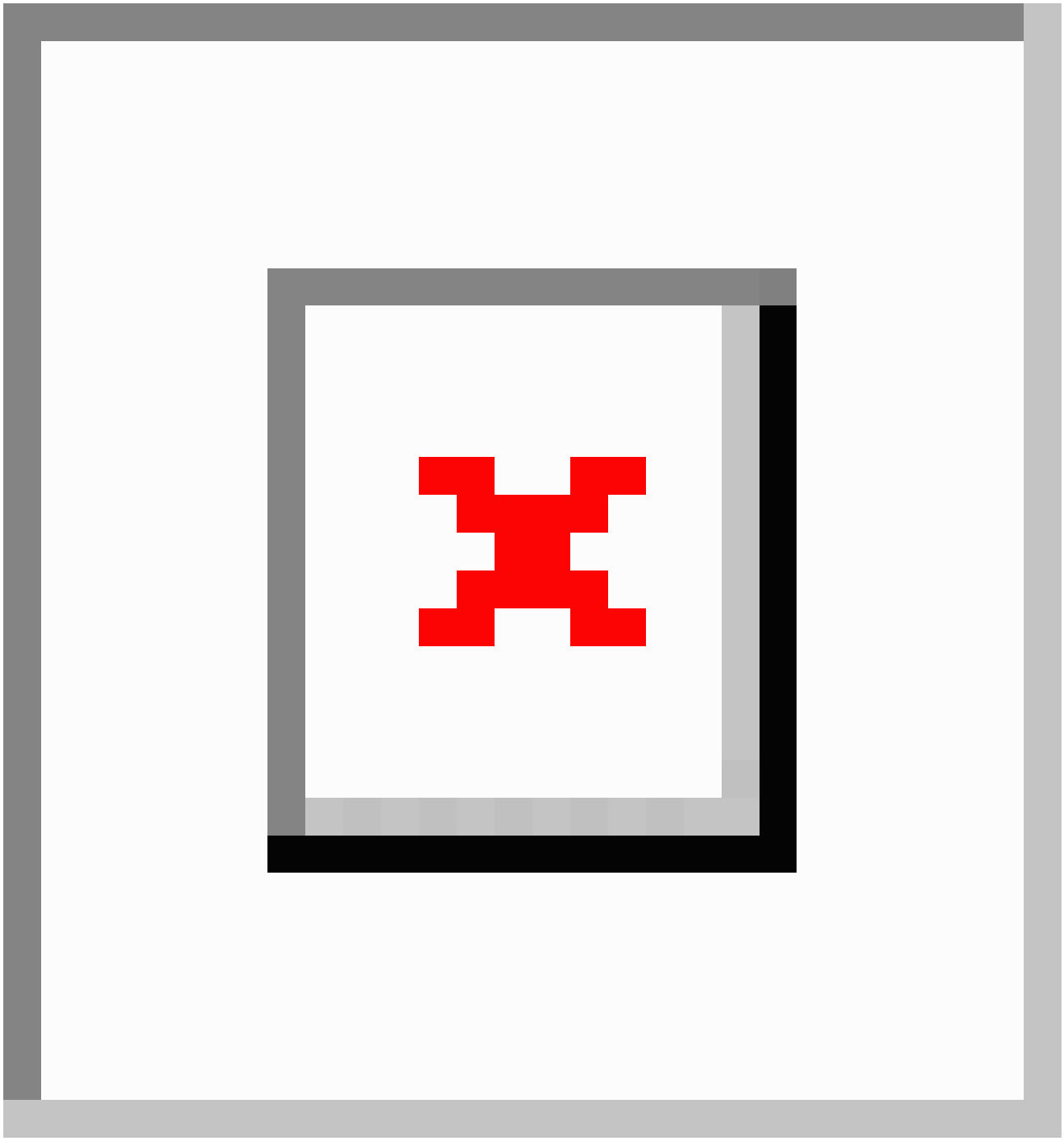
The focal point of the intervention geared toward implementing the NBDW entailed orchestrating the reorganization across the involved hospital research entities. This restructuring initiative encompassed not only the resolution of technical hurdles but also the delineation of governance structures, regulatory frameworks, and parameters governing data access.

This implementation report adheres to the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) reporting guidelines [26].

From Blueprint to the Functional Organization of the NBDW

Implementation of the NBDW required the de novo organization of a functional research network within a 3-fold structure (Figure 1). First, enforcement of the governance policy and NBDW legal framework was achieved by the Research Administration Department (RAD), which assumed legal and ethical responsibility. Second, the creation of a clinical data center (CDC) took charge of the scientific question and the potential need for NBDW data to coassume scientific responsibility with project investigators. The last layer of technical responsibility was established by the hospital's Information Technology Department (ITD). IT priorities are defined by legal and governance constraints, and address the data needs of each project. The new organization allows the distribution of tasks and responsibilities in the implementation and management of the NBDW.

Figure 1. A functional framework dedicated to the CHUN NBDW. A 3-fold governance structure constitutes the functional research framework accompanying the development of the NBDW. The ITD ensures the IT infrastructure's quality, transparency, and data deidentification. Through an ETL process, data flows are extracted and loaded to the NBDW. The RAD is an administrative department that assumes legal and ethical responsibility, and lays the regulatory framework that will define data use and access policies. The CDC assumes scientific data stewardship, is in charge of scientific responsibility, supports the methodology of the study, and grants data access. Those three structures are organized and structured to ensure research quality, ethics, and technical transparency. White arrows represent the data request and access process undertaken by investigators. CDC: clinical data center; CHUN: Centre Hospitalo-Universitaire de Nantes; ETL: extract, transform, and load; ITD: Information Technology Department; NBDW: Nantes Biomedical Data Warehouse; NLP: natural language processing; RAD: Research Administration Department.



Legal Functional Framework

On behalf of the hospital, the RAD defines the framework of data use and access policies, security, and patient consent and privacy, and is in charge of the construction, organization, and enforcement of the NBDW regulatory framework. It works in conjunction with the hospital data protection officer, represented by an officer specifically dedicated to research data. Furthermore, the RAD structure requires documented monitoring to ensure a constant adaptation of the NBDW regulatory framework to answer to changes in legal standards. All projects are publicly described on the web [27] for all potential patients contributing data.

IT Stewardship and Interoperability

The ITD is in charge of the collection, storage aggregation, quality, and integrity of clinical data. This structure carries out a continuous process to conform to technical standards. Setting up the addition of new hospital data sources is an ongoing process that started in July 2018 and is still relevant.

The NBDW is set up on virtual machines. An Oracle database enables massive data storage, integrating multiple hospital data sources into a unique set of tables containing data on patients, consultations, diagnoses, laboratory results, medical notes, and inpatient drug administration. The NBDW uses eHOP [22,28] software to organize and query the database. eHOP is a platform developed by Rennes University using a public-private partnership on the enterprise application integration Enovacom Suite Version 2 (ESV2) of Enovacom (Orange Business Service-Santé). eHOP carries out the acquisition, transformation, and integration of the data coming from various HISs with different formats and standards (Health Level 7 [HL7]; Harmoniser et promouvoir l'informatique médicale [HPRIM]; PN13; Logical Observation Identifiers Names and Codes [LOINC]; and Word documents, PDF, CSV, and text). ESV2 provides an automatic, scheduled (daily, weekly, or monthly), and monitored data supply from the NBDW.

Scientific Data Stewardship and Mediated Access to BDW

Beginning in 2018, the CDC has had a specific team dedicated to the reuse of health data for research, relying on the expertise of public health doctors, data and computer scientists, epidemiologists, biostatisticians, and project managers promoting epidemiology analyses and providing support to clinical investigators and researchers in data access. In conjunction with the ITD, the CDC defines the standards necessary for data integration and data use practices, and ensures data control and the scientific quality of the analyses.

Target and Data Access

The end users of the NBDW are investigators (internal or external to CHUN) aiming to advance the institution's research. A per-project access is created for CHUN investigators through a standardized approval process consisting of 4 main steps (Figure 1). First, investigators register their request and submit a research protocol through a portal hosted by the CHUN intranet. Second, the CDC validation board, which meets once a week, analyzes three dimensions of the submitted research projects: compliance with ethical principles, scientific relevance, and feasibility. On completion of the approval process, the project is registered on an internal database for the completion of legal requirements. Third, the CDC processes all the necessary queries on the NBDW to select a group of patients relevant to the scientific question. During this step, ongoing collaboration with the investigator is necessary, in particular, to precisely define the scientific purpose, eligible population, and data of interest. Fourth, data are made available internally through a data mart, hosted by the CHUN intranet. The data mart system provides regulated, parsimonious (only the required data regarding the targeted population), and time-limited access to investigators. Data are completely deidentified. Moreover, investigators can make simple queries and seek data on patients contained in the data marts through eHOP, which is enriched with a set of tools for simple textual and structured data queries.

Projects requiring data management and extraction to integrate a research database are declared to the public registry of CHUN projects as a guarantee of transparency and to allow patient opposition. At this step, more complex methods for the extraction of information through natural language processing (NLP) [29], regular expression tools, or other structured data [30] may be applied. Finally, data extraction is constrained to strictly necessary data, following the parsimony principle, and only if access to data can be done in a secure environment.

In the case of a project supported by an external project leader from CHUN (academic or private partner), the same process as described above takes place with the exception of the following differences: the project might be supported by a clinical team that submits the research protocol through the portal; a partnership agreement must be signed between the hospital and the partner, and the data mart is only available through a specific virtual working space (data are still internally hosted).

Data Protection and Patient Consent

To comply with national and international privacy regulations, data integration is subjected to a deidentification algorithm. Data are stored in two independent and separate Oracle schemas to separate pseudonymized data from nominative or other directly reidentifying information to which access is strongly limited. Data separation is supplemented by access management and traceability of the actions carried out (ie, AuditLog). Most notably, the platform includes a functionality for collecting and applying patient consent to the use of personal data, ensuring compliance with French law and European General Data Protection Regulation requirements [31].

Regulatory Approval

In alignment with the French Data Protection Act (Loi Informatique et Libertés, 1978), the use of personal data for health research and evaluation requires compliance with a reference methodology, representing good practices. Without such compliance, personal data use must be authorized by the Commission Nationale de l'Informatique et des Libertés (CNIL; French National Commission for Information Technology and Civil Liberties). At the launch of the NBDW, no research methodology existed for data warehouses in the field. Therefore, approval from the CNIL was mandatory to initiate implementation. Submission to the CNIL covered legal responsibilities, data processing details, access, governance, and more. Comprehensive data access details were provided, extending to researchers whether affiliated with CHUN or not. Private entities are permitted to engage in research projects based on the NBDW, ensuring adherence to this resolution and French regulations. The Data Protection Impact Assessment for NBDW was an integral part of the submission to the CNIL, serving as a mandatory document. The authorization to set up and use the NBDW was granted on July 19, 2018, by the CNIL (resolution 2018 - 295).

Budget Planning and Sustainability

Estimating the costs associated with implementing and maintaining a data warehouse is challenging owing to several factors. First, the NBDW is part of an institutional strategy, making it difficult to consider it as a stand-alone entity. Second, implementing it involves the collaboration and coordination of multiple structures and experts, complicating the estimation of resource use. Third, hidden costs are difficult to anticipate and consider, including system failures and delays, unplanned license renewals and upgrades, adjustments to regulatory and legal requirements, unplanned changes in HISs, and infrastructure upgrades. We made an estimation by considering three budget lines—infrastructure, license, and human resources—and two different periods—completion in 5 years (2018 - 2022) and maintenance in 2 years (2022 and 2023)—for a total of €2.6 million (US $2.8 million).

In terms of sustainability, an annual operational budget is allocated for maintenance and updates. Specific needs for the integration of new hospital data sources are financed through project-based funding. Moreover, an economic model is currently being defined to incorporate additional charges for infrastructure costs in the case of external research projects.

Ethical Considerations

An ethics statement is included in the regulatory approval granted by CNIL with resolution number 2018 - 295 [32].

Implementation (Results)

Description of the Sources, Concepts, and Contents

The NBDW integrates multiple hospital data sources into a unique and structured set of tables containing data on patient demographic and administrative records, inpatient drug administration, inpatient constants and anthropometric scores and metrics, anatomic pathology notes, inpatient and outpatient medical laboratory results, narrative medical notes (including admission/discharge summaries, inpatient anesthesia notes, outpatient consultation notes, nurse notes), *International Classification of Diseases, 10th Revision* (*ICD-10*) and French Classification Commune des Actes Médicaux (CCAM; Common Classification of Medical Procedures) codes for inpatient diagnoses and procedures, and medical imaging reports. Table 1 shows principal concepts and contents according to different HIS data sources or software integrated up to now. Some data sources contain only narrative notes, and some sources contain both unstructured and structured data.

Table . Nantes University Hospital Biomedical Data Warehouse principal sources and concepts. Data extracted May 10, 2023.

Concepts	Software	Period	Patients, n	Documents, n	Documents in 2022, n	Structured data, n
Inpatient drug prescriptions	MILLENNIUM	2015-today	318,456	39,681,513	5,673,000	248,289,146
Cardiology narrative notes	CARDIOREPORT	2015-today	9278	38,041	6644	—[a]
Consultation clinical narrative notes	GAM-CLINICOM	2002-today	1,053,386	6,507,860	22,982	—
Constants and anthropometric data	MILLENNIUM	2015-today	440,250	3,306,589	638,969	61,142,757
Anatomic pathology notes	DIAMIC	2015-today	131,812	229,081	27,244	1246
Biology laboratory results	DXLAB	2012-today	701,804	10,752,384	1,672,218	155,825,060
Clinical narrative notes	MILLENNIUM	2015-today	569,114	3,967,073	907,573	5,395,233
ICD-10[b] and clinical procedure codes	CLINICOM	2006-today	725,802	5,318,712	401,607	105,246,620
Radiology reports	QDOC	2015-today	284,937	904,625	122,900	—
Nurse transmissions	TRANSMISSIONS	2017-today	131,508	1,546,114	320,293	1,546,114

[a]Not applicable.

[b]*ICD-10*: *International Classification of Diseases, 10th Revision.*

NBDW Figures and Populations

CHUN, a tertiary care hospital ranked seventh in France in terms of activity [33], provides care over a population catchment area of 1.4 million inhabitants. It provides follow-up and long-term health care for both in- and outpatients. It has 2993 hospital beds, delivers 4380 babies, and conducts more than 1 million consultations and external medical procedures per year [34]. It also carries out practical teaching for 1200 medical students, 800 medical residents, and over 2000 non–medical students.

The NDBW includes information on approximately 1.5 million patients admitted between 2003 and 2022 (Table 2). More than 1.2 million hospitalizations are associated with approximately

12.3 million *ICD-10*–coded diagnoses and 7.3 million clinical procedure codes. Together with more than 6.3 million external consultations, the NBDW contains more than 11 million textual documents. These narrative notes integrated as free-text documents can be interrogated and turned into structured data for research.

The yearly number of patients, hospitalizations, consultations, and narrative notes has increased over time (Multimedia Appendix 1) with growth rates between 2003 and 2019 ranging from 109% (hospitalizations) to 430% (outpatient consultations).

Table . Nantes University Hospital Biomedical Data Warehouse figures and contents, 2003 - 2022.

Contents	Since 2003, n	2022 only, n
Patients[a]	1,597,498	300,804
Hospitalizations[b]	2,635,809	183,361
Outpatient consultations	6,358,271	524,948
Clinical narrative notes[c]	11,634,761	826,670
Diagnoses[d]	12,251,148	956,688
Clinical procedures[e]	7,272,346	504,571

[a]Patients with ≥1 clinical narrative note or a structured document, including inpatient and patients admitted for outward consultations.

[b]Inpatient hospitalizations in medical, surgical, and obstetric services, including complete hospitalizations, day-hospital admissions, and recurring visits.

[c]Clinical narrative notes (with the exclusion of vital signs and anthropometric data; *International Classification of Diseases, 10th Revision* [*ICD-10*] and clinical procedure codes; laboratory results; inpatient drug administrations; and nurse transmissions).

[d]Medical diagnoses following the *ICD-10* for medical, surgical, and obstetric hospitalizations.

[e]Classification Commune des Actes Médicaux for medical, surgical, and obstetric hospitalizations.

Projects and Effectiveness Outcomes

The availability of NBDW data makes it possible to provide data in response to a wide array of scientific questions and the need for data in the analysis and management of care and organization. Prior to the creation of NBDW data, researchers were limited to the interrogation of medico-administrative–structured information out of the scope of data reuse consent. It only covered structured data such as *ICD-10* diagnoses associated with hospitalizations; medical procedures through CCAM codes; and, to a lesser extent owing to availability issues, laboratory results. Obtaining results from different queries performed independently on different HISs and data manager services was time-consuming if not impossible for both legal and technical reasons.

The NBDW currently facilitates queries of clinical concepts in both structured and unstructured free-text notes in an integrated environment and with respect to data protection policies and laws. BDW data requests may be divided into three different types of research projects according to their purpose: (1) optimize patient screening in both clinical trials and observational studies, (2) enrich case reports or electronic case report forms for disease surveillance, and (3) evaluate and improve clinical practices and resource management. To illustrate different types of studies, three concrete examples of data use are described in Multimedia Appendix 2 [35-37].

The first outcome to measure the effectiveness of the NBDW was defined as the number of studies supported through the project tracking portal (Figure 1). Since 2018, 577 requests have been made and treated by the CDC. Among them, 269 projects involved patients included thanks to NBDW queries and research tools (second outcome), and for 115 of them, data marts were created to give investigators access to data (third outcome).

Discussion

Lessons Learned

The development of clinical data warehouses has provided unprecedented access to a large amount of diverse data from clinical care. However, it requires a dedicated effort in terms of governance, data access rules with respect to patient consent and data protection, and technical challenges. The reorganization of structures within a functional research framework is the first factor in the success of the NBDW. Collaboration between departments has not only facilitated seamless communication but also engendered the innovation necessary to deal with the complexities of health care data management. It was an opportunity to test new IT technologies such as distributed infrastructure and anonymization techniques such as deidentification. The interaction between structures has also been a fundamental element in the process of obtaining BDW authorization from the CNIL. Indeed, the obtention of regulatory approval is the result of a long negotiation that required expertise in addressing legal, technical, and scientific research requirements (see Regulatory Approval section for more details).

The creation of the CDC, the result of multidisciplinary teamwork composed of computer scientists, NLP engineers, statisticians, physicians, epidemiologists, and project managers, has probably been the second factor of success. The weekly CDC validation boards, supported by the RAD and ITD structures, verify project compliance vis-à-vis three aspects: ethics, scientific relevance, and practical feasibility, giving support and access to the NBDW in a secure context. However, it is also a leading driver in the shift toward data-driven governance in hospitals.

Implementation of the NBDW has required a continuous and still relevant process to conform to new regulatory and technical standards, and to add new hospital sources and ongoing improvements. An important lesson learned was that each of the 269 NBDW projects in the past 5 years has been an opportunity to revisit the contents of the BDW, further the quality control process, and lead the data transformation process, creating data use value.

Establishing networks and working together is probably the best lesson learned. In France, some experiences have led to changes in health data foresight [28] and promoted the implementation of interregional [22] and national hubs. The NBDW has benefited from and contributed to the "Ouest Data Hub," a network of Western France university hospital data warehouses. The aim is both to facilitate the reproducibility of data analyses, share resources and best querying practices, and promote adapted and standardized terminologies and nomenclatures between centers. Moreover, BDWs use the same software for both integration and querying, allowing them to be interrogated using consistent queries and rules. This approach ensures a high level of interoperability and accessibility, facilitating seamless interaction and adherence to the Findable, Accessible, Interoperable, Reusable (FAIR) principles.

In hindsight, if this process were to be revisited, there are certain facets that we would address differently. Specifically, in the initial stages of implementing the NBDW, primary emphasis was placed on deploying the software and IT infrastructure recommended by the regional network, ODH, aimed at facilitating the establishment of the repository and its querying system. While acknowledging the benefits inherent in this strategy, a more thorough examination of alternative IT solutions is warranted to mitigate reliance on a singular approach and to streamline potential transitions to alternative and variated options.

Conclusions

In conclusion, conducting health studies using electronic health records requires careful attention to ensure accurate results owing to a lack of a systematic quality process. The data quality control procedure is a long and necessary process [17,38,39]. The future challenge will be the setting up of standardized and shared quality control pipelines to ensure quality results, not only at the local level but also in a regional and national context of future data sharing. By extending long-term investments in IT and data in care institutions, the development of NLP and text-mining tools will further accelerate the use of BDW, facilitating data-driven decision-making discussions from top management to patients.

Any research project and analysis of care-based research performed in health management institutions could benefit from the deployment of organized data access. The collaborative nature of data production and the information and privacy protection for patients require mediated and expert access to the BDW. It demonstrates that technical solutions are partial answers to better data-driven practices and must lead to a clear governance strategy

Acknowledgments

The authors wish to thank P Boistard, M Lebigre, C Cartau, A Magnan, P Sudreau, P Lecerf, Dr S Sacher-Huvelin, Dr V Guardiolle, Dr J Esbelin, M Lazarevic, A Royer, and AC de Reboul for their assistance.

This work was financially supported, in part, by the Agence Nationale de la Recherche AIBy4 under contract ANR-20-THIA-0011 and the cluster DELPHI - NExT under contract ANR-16-IDEX-0007, and integrated into the France 2030 plan by Région Pays de la Loire and Nantes Métropoles.

Conflicts of Interest

PAG is the founder of Methodomics (2008) and the cofounder of Big data Santé (2018). He consults for major pharmaceutical companies and start-ups, all of which are handled through academic pipelines (AstraZeneca, Biogen, Boston Scientific, Cook, Docaposte, Edimark, Ellipses, Elsevier, Methodomics, Merck, Mérieux, Octopize, and Sanofi-Genzyme). PAG is a volunteer board member at the AXA not-for-profit mutual insurance company (2021). He has no prescription activity with either drugs or devices. The other authors declare no potential conflicts of interest to disclose.

Multimedia Appendix 1
Nantes University Hospital Biomedical Data Warehouse yearly data volume by type of data.
[DOCX File, 131 KB - medinform_v12i1e50194_app1.docx]

Multimedia Appendix 2
Three projects as an example of case experiences based on the Nantes University Hospital Biomedical Data Warehouse.
[DOCX File, 232 KB - medinform_v12i1e50194_app2.docx]

Checklist 1
iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations).
[DOCX File, 29 KB - medinform_v12i1e50194_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. Safran C, Bloomrosen M, Hammond WE, et al. 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. [doi: 10.1197/jamia.M2273] [Medline: 17077452]
3. Haarbrandt B, Tute E, Marschollek M. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. J Biomed Inform 2016 Oct;63:277-294. [doi: 10.1016/j.jbi.2016.08.007] [Medline: 27507090]
4. 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. [doi: 10.2196/17687] [Medline: 32852280]
5. Chute CG, Beck SA, Fisk TB, Mohr DN. The enterprise data trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc 2010;17(2):131-135. [doi: 10.1136/jamia.2009.002691] [Medline: 20190054]
6. Pavlenko E, Strech D, Langhof H. Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. BMC Med Inform Decis Mak 2020 Jul 11;20(1):157. [doi: 10.1186/s12911-020-01177-z] [Medline: 32652989]
7. Holmes JH, Elliott TE, Brown JS, 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;21(4):730-736. [doi: 10.1136/amiajnl-2013-002370] [Medline: 24682495]
8. Bloomrosen M, Detmer D. Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association. J Am Med Inform Assoc 2008;15(6):715-722. [doi: 10.1197/jamia.M2905] [Medline: 18755988]
9. Rosenbaum S. Data governance and stewardship: designing data stewardship entities and advancing data access. Health Serv Res 2010 Oct;45(5 Pt 2):1442-1455. [doi: 10.1111/j.1475-6773.2010.01140.x] [Medline: 21054365]
10. 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]
11. Fennelly O, Cunningham C, Grogan L, et al. Successfully implementing a national electronic health record: a rapid umbrella review. Int J Med Inform 2020 Dec;144:104281. [doi: 10.1016/j.ijmedinf.2020.104281] [Medline: 33017724]
12. Constant Dit Beaufils P, Karakachoff M, Gourraud PA, Bourcier R. Management of unruptured intracranial aneurysms: how real-world evidence can help to lift off barriers. J Neuroradiol 2023 Mar;50(2):206-208. [doi: 10.1016/j.neurad.2023.01.156] [Medline: 36724868]
13. Kurian AW, Mitani A, Desai M, et al. Breast cancer treatment across health care systems: linking electronic medical records and state registry data to enable outcomes research. Cancer 2014 Jan 1;120(1):103-111. [doi: 10.1002/cncr.28395] [Medline: 24101577]
14. Greenberg AE, Hays H, Castel AD, et al. Development of a large urban longitudinal HIV clinical cohort using a web-based platform to merge electronically and manually abstracted data from disparate medical record systems: technical challenges and innovative solutions. J Am Med Inform Assoc 2016 May;23(3):635-643. [doi: 10.1093/jamia/ocv176] [Medline: 26721732]
15. 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. [doi: 10.1016/j.ijmedinf.2019.05.018] [Medline: 31445247]
16. Artemova S, Madiot PE, Caporossi A, PREDIMED group, Mossuz P, Moreau-Gaudry A. PREDIMED: clinical data warehouse of Grenoble Alpes University Hospital. Stud Health Technol Inform 2019 Aug 21;264:1421-1422. [doi: 10.3233/SHTI190464] [Medline: 31438161]
17. Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: a 8-years follow-up experience. Int J Med Inform 2017 Jun;102:21-28. [doi: 10.1016/j.ijmedinf.2017.02.006] [Medline: 28495345]
18. Wack M. Installation d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy: déploiement technique, intégration et gouvernance des données [Doctoral thesis].: Université de Lorraine; 2017 Oct 7 URL: http://docnum.univ-lorraine.fr/public/BUMED_T_2017_WACK_MAXIME.pdf [accessed 2024-06-17]
19. Pressat-Laffouilhère T, Balayé P, Dahamna B, 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 8;22(1):34. [doi: 10.1186/s12911-022-01762-4] [Medline: 35135538]
20. Entrepôts de données de santé hospitaliers en France. Haute Autorité de Santé. 2022 Nov 17. URL: https://www.has-sante.fr/jcms/p_3386123/fr/entrepots-de-donnees-de-sante-hospitaliers-en-france [accessed 2023-01-03]
21. Doutreligne M, Degremont A, Jachiet PA, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: a case study in France. PLOS Digit Health 2023 Jul 6;2(7):e0000298. [doi: 10.1371/journal.pdig.0000298] [Medline: 37410797]

22. Madec J, Bouzillé G, Riou C, et al. eHOP clinical data warehouse: from a prototype to the creation of an inter-regional clinical data centers network. Stud Health Technol Inform 2019 Aug 21;264:1536-1537. [doi: 10.3233/SHTI190522] [Medline: 31438219]
23. Bocquet F, Raimbourg J, Bigot F, Simmet V, Campone M, Frenel JS. Opportunities and obstacles to the development of health data warehouses in hospitals in France: the recent experience of comprehensive cancer centers. Int J Environ Res Public Health 2023 Jan 16;20(2):1645. [doi: 10.3390/ijerph20021645] [Medline: 36674399]
24. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics initiatives: two national projects to promote data sharing in healthcare. Yearb Med Inform 2019 Aug;28(1):195-202. [doi: 10.1055/s-0039-1677917] [Medline: 31419832]
25. Goldberg M, Zins M. Health data hub: why and how? Med Sci (Paris) 2021 Mar;37(3):271-276. [doi: 10.1051/medsci/2021016] [Medline: 33739275]
26. Perrin Franck C, Babington-Ashaye A, Dietrich D, et al. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. J Med Internet Res 2023 May 10;25(1):e46694. [doi: 10.2196/46694] [Medline: 37163336]
27. Liste de recherches sur données et/ou échantillons menées au CHU de Nantes, notamment à partir de l'entrepôt. Centre Hospitalier Universitaire de Nantes. URL: https://www.chu-nantes.fr/liste-des-etudes-menees-au-chu-de-nantes-utilisant-des-donnees-de-l-entrepot-de-recherche [accessed 2024-01-26]
28. Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. Stud Health Technol Inform 2015;210:702-706. [Medline: 25991243]
29. Labrak Y, Bazoge A, Dufour R, et al. DrBERT: a robust pre-trained model in French for biomedical and clinical domains. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): Association for Computational Linguistics; 2023:16207-16221. [doi: 10.18653/v1/2023.acl-long.896]
30. Guardiolle V, Bazoge A, Morin E, et al. Linking biomedical data warehouse records with the national mortality database in France: large-scale matching algorithm. JMIR Med Inform 2022 Nov 1;10(11):e36711. [doi: 10.2196/36711] [Medline: 36318244]
31. General Data Protection Regulation (GDPR). URL: https://gdpr-info.eu/ [accessed 2023-06-06]
32. Délibération 2018-295 Du 19 Juillet 2018. Légifrance. 2018 Oct 23. URL: https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000037509951 [accessed 2024-06-11]
33. Bases statistiques SAE. Données statistiques publiques en santé et social. 2015 Jan 6. URL: https://data.drees.solidarites-sante.gouv.fr/explore/dataset/708_bases-statistiques-sae/ [accessed 2021-12-28]
34. En chiffres. Centre Hospitalier Universitaire de Nantes. URL: https://www.chu-nantes.fr/activite-et-chiffres-cles [accessed 2023-01-03]
35. Bourcier R, Chatel S, Bourcereau E, et al. Understanding the pathophysiology of intracranial aneurysm: the ICAN Project. Neurosurgery 2017 Apr 1;80(4):621-626. [doi: 10.1093/neuros/nyw135] [Medline: 28362927]
36. Cotton F, Kremer S, Hannoun S, Vukusic S, Dousset V, Imaging Working Group of the Observatoire Français de la Sclérose en Plaques. OFSEP, a nationwide cohort of people with multiple sclerosis: consensus minimal MRI protocol. J Neuroradiol 2015 Jun;42(3):133-140. [doi: 10.1016/j.neurad.2014.12.001] [Medline: 25660217]
37. Lucas DN, Yentis SM, Kinsella SM, et al. Urgency of caesarean section: a new classification. J R Soc Med 2000 Jul;93(7):346-350. [doi: 10.1177/014107680009300703] [Medline: 10928020]
38. Jantzen R, Rance B, Katsahian S, Burgun A, Looten V. The need of an open data quality policy: the case of the "transparency - health" database in the prevention of conflict of interest. Stud Health Technol Inform 2018;247:611-615. [Medline: 29678033]
39. Looten V, Kong Win Chang L, Neuraz A, et al. What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse. Comput Methods Programs Biomed 2019 Nov;181:104825. [doi: 10.1016/j.cmpb.2018.12.030] [Medline: 30612785]

Abbreviations

BDW: biomedical data warehouse
CCAM: Classification Commune des Actes Médicaux (English: Common Classification of Medical Procedures)
CDC: clinical data center
CHUN: Centre Hospitalo-Universitaire de Nantes (English: University Hospital of Nantes)
CNIL: Commission Nationale de l'Informatique et des Libertés (English: French National Commission for Information Technology and Civil Liberties)
ESV2: Enovacom Suite Version 2
FAIR: Findable, Accessible, Interoperable, Reusable
HIS: hospital information system
HL7: Health Level 7
HPRIM: Harmoniser et promouvoir l'informatique médicale

***ICD-10*:** *International Classification of Diseases, 10th Revision*
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
ITD: Information Technology Department
LOINC: Logical Observation Identifiers Names and Codes
NBDW: Nantes University Hospital Biomedical Data Warehouse
NLP: natural language processing
RAD: Research Administration Department

Edited by C Perrin; submitted 22.06.23; peer-reviewed by D Reuter, K Gierend, O Steichen; revised version received 08.04.24; accepted 17.04.24; published 24.06.24.

Please cite as:
Karakachoff M, Goronflot T, Coudol S, Toublant D, Bazoge A, Constant Dit Beaufils P, Varey E, Leux C, Mauduit N, Wargny M, Gourraud PA
Implementing a Biomedical Data Warehouse From Blueprint to Bedside in a Regional French University Hospital Setting: Unveiling Processes, Overcoming Challenges, and Extracting Clinical Insight
JMIR Med Inform 2024;12:e50194
URL: https://medinform.jmir.org/2024/1/e50194
doi:10.2196/50194

Implementation Report

Design and Implementation of an Inpatient Fall Risk Management Information System

Ying Wang[1,2], MSM; Mengyao Jiang[2], MSN; Mei He[2], MSN; Meijie Du[2], MSN

[1]School of Management, Wuhan University of Technology, Wuhan, China

[2]Department of Nursing, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Corresponding Author:
Ying Wang, MSM
School of Management
Wuhan University of Technology
122 Luoshi Road
Hongshan District
Wuhan, 430070
China
Phone: 86 027 83662317
Email: wangying_tjh@hotmail.com

Abstract

Background: Falls had been identified as one of the nursing-sensitive indicators for nursing care in hospitals. With technological progress, health information systems make it possible for health care professionals to manage patient care better. However, there is a dearth of research on health information systems used to manage inpatient falls.

Objective: This study aimed to design and implement a novel hospital-based fall risk management information system (FRMIS) to prevent inpatient falls and improve nursing quality.

Methods: This implementation was conducted at a large academic medical center in central China. We established a nurse-led multidisciplinary fall prevention team in January 2016. The hospital's fall risk management problems were summarized by interviewing fall-related stakeholders, observing fall prevention workflow and post–fall care process, and investigating patients' satisfaction. The FRMIS was developed using an iterative design process, involving collaboration among health care professionals, software developers, and system architects. We used process indicators and outcome indicators to evaluate the implementation effect.

Results: The FRMIS includes a fall risk assessment platform, a fall risk warning platform, a fall preventive strategies platform, fall incident reporting, and a tracking improvement platform. Since the implementation of the FRMIS, the inpatient fall rate was significantly lower than that before implementation (P<.05). In addition, the percentage of major fall-related injuries was significantly lower than that before implementation. The implementation rate of fall-related process indicators and the reporting rate of high risk of falls were significantly different before and after system implementation (P<.05).

Conclusions: The FRMIS provides support to nursing staff in preventing falls among hospitalized patients while facilitating process control for nursing managers.

(JMIR Med Inform 2024;12:e46501) doi:10.2196/46501

KEYWORDS

fall; hospital information system; patient safety; quality improvement; management; implementation

Introduction

Context

Falls are one of the nursing-sensitive indicators for nursing care [1], which are a leading cause of fatal and nonfatal health loss globally [2,3]. Reducing and preventing falls has become an international health priority. Falls—common adverse events reported in hospitals—have been identified as a nursing-sensitive quality indicator of patient care.

Given the growing technological progress, health IT may help enhance the quality and safety of provided care, facilitating the effectiveness and efficiency of the clinical workflow, and supporting the provision of integrated multidisciplinary care [4-11]. The hospital information system (HIS) is a promising

approach to improve care quality and safety in the complex hospital environment. Despite extensive literature on fall risk factors and preventive strategies [12-18], few studies have focused on health information systems for managing inpatient falls.

Problem Statement

To address these issues, we formed a nurse-led multidisciplinary fall prevention team in January 2016, including the hospital administrative staff, quality management specialists, physicians, nurses, pharmacists, and informatics staff. This team retrospectively analyzed 19 inpatient fall cases that occurred in 2015 (fall rate 0.015‰), ranking first among all in-hospital nursing adverse events. Among the fall cases, 30%-40% of patients had grade ≥3 injuries, which significantly exceeded the 3.978% proportion seen in similar hospitals during this period. Falls caused severe harm and financial burden to inpatients, with 3 patients experiencing severe head injuries and 2 having hip fractures. The longest hospital stay resulting from falls reached 36 days.

The hospital's fall risk management problems were summarized by interviewing fall-related stakeholders, observing fall prevention workflow and postfall care process, and investigating patients' satisfaction; these included (1) nonachievement of real-time fall risk assessment, real-time uploading, and information sharing; (2) absence of fall risk warning management; (3) complicated fall risk management workflow; (4) absence of process control in fall prevention (such as process control for different fall risk levels, process control for different time nodes, etc); and (5) lack of standardized pathways for inpatient fall incident reporting and improvement tracking.

Similar Interventions

Several studies have highlighted the benefits of using health information systems for patient fall management. For example, Giles et al [19] reported that data collected from nursing information systems can be used to identify high-fall-risk patients. Mei et al [20] designed an electronic patient fall reporting system in a US long-term residential care facility, which could improve the fall reporting process and subsequent quality improvement efforts. Katsulis et al [21] combined the Fall TIPS (Tailoring Interventions for Patient Safety) [22] with a clinical decision support system, which increased its ease of use over the paper version. Jacobsohn et al [23] developed an automated clinical decision support system for identifying and referring older adult emergency department patients at the risk of future falls. Mlaver et al [24] at the Brigham and Women's hospital developed a valuable electronic health record–embedded dashboard that collected inpatient fall risk data. However, the abovementioned fall information system only focused on a specific domain of fall management. So far, there is still no report about an HIS for overall fall risk management.

Aims and Objectives

This implementation aims to design and implement a fall risk management information system (FRMIS) to reduce falls among inpatients and improve nursing quality. Our goal is to create a culture of safety and reduce the incidence of falls hospital-wide, ensuring the well-being and security of all patients.

Methods

This study adhered to the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) checklist [25].

Blueprint Summary

This FRMIS consists of 4 major functional platforms to facilitate comprehensive fall prevention pathway management as shown in Textbox 1.

Textbox 1. The 4 major functional platforms of the fall risk management information system.

A fall risk assessment platform (Multimedia Appendix 1):

The assigned nurse uses a personal digital assistant to conduct fall risk assessments within 4 hours of patient admission. Upon completion, the personal digital assistant automatically compiles the Morse Fall Risk Score [26,27] and risk level, marking it in the electronic nursing record. All patients' Morse Fall Risk Scores are collected and shared in real time through the information platform. Simultaneously, nurses receive nursing guidelines specific to different fall risk levels. They implement corresponding fall prevention measures such as hanging “Fall Prevention” warning signs near high-risk patients' beds, distributing “Fall Prevention Information Sheets” to guide patients and their families on preventive measures, and documenting and passing on relevant information during shift changes. The head nurse conducts daily inspections and guidance on the accuracy of Morse fall assessments and the implementation of fall prevention measures, upon completion of departmental reviews.

A fall risk warning platform (Multimedia Appendix 1):

Patients at different fall risk levels are color-coded for easy identification: high-risk (Morse Fall Risk Score≥45), red; moderate (score approximately 25-44), yellow; and low (score approximately 0-24), green. The fall risk warning module comprehensively displays the daily number of high-risk falls, department distribution and ranking, percentage of the population at the risk of falls, specific bed locations, medical diagnoses, Morse Fall Risk Scores, and assessment times through charts and color-coded indicators. This provides nursing managers real-time insights into the key populations, departments, and information related to fall risk management, enabling proactive fall risk prevention and providing precise information support for effective fall prevention process control.

A fall preventive strategies platform (see Multimedia Appendix 1):

Evidence-based fall prevention strategies are developed, incorporating fall event analysis and expert discussions to extract key process monitoring indicators. An electronic fall prevention bundle strategies quality tracking checklist was established for accurate assessment of fall risk, increased awareness of preventive measures, enhanced handover process for high-risk patients, environmental safety, implementation of fall prevention knowledge training, and guidance on proper use of assistive devices. Nursing department and ward-level managers can use mobile devices (iPads) to conduct targeted goal management and quality inspections of fall prevention strategies. Real-time monitoring is conducted on key fall process indicators such as accuracy of Morse fall risk assessments, implementation of health education, adherence to handover procedures, and compliance with environmental safety measures.

A fall incident reporting and tracking platform (see Multimedia Appendix 1):

The platform regulates the reporting process for inpatient fall events. After a fall incident occurs, the ward head nurse promptly logs into the fall incident reporting platform to proactively report the incident. They provide details such as the time and location of the fall, the sequence of events, whether the patient was injured, the extent of the injury, and the emergency treatment process. Once the information platform receives the ward's report, it immediately sends text messages to the chief nurse and members of the nursing department's safety management team. On the platform, safety management team members can quickly trace the Morse Fall Risk Score, risk level, appropriateness of fall prevention interventions, timeliness of assessments, and any dynamic evaluations associated with that patient. After gaining a comprehensive understanding of the patient's relevant information, they visit the ward in a timely manner to conduct on-site inspections and tracking. They provide guidance to the department by applying root cause analysis to thoroughly analyze the fall event, identify the underlying causes, and propose areas of improvement directly on the web-based platform. Ward head nurses and the chief nurse can access expert guidance instantly on the platform and make necessary improvements based on the advice provided.

Technical Design

The FRMIS was developed using an iterative design process, involving collaboration among health care professionals, software developers, and system architects. The design aimed to create a user-friendly interface, incorporate data integration capabilities, and enable real-time reporting functionalities. In order to meet the usage needs of both PC and mobile devices, the development language selected for this system includes C#, jQuery, and Java; the development tools used were Visual Studio (Microsoft Corp) and Eclipse (The Eclipse Foundation), and the development platforms used were Windows and Android.

Target

The FRMIS was designed to assist nursing staff in preventing inpatient falls through IT, facilitating process control for nursing managers and ensuring patient safety.

Data

Our hospital has a dedicated computer center, which serves as the technical support department for network security. It is responsible for the construction and operation of hospital network security protection measures. The collection of various data in the FRMIS complies with relevant national laws and regulations. The data collection scope follows the principle of “minimum necessary” and adopts measures such as data desensitization, data encryption, and link encryption to prevent data leakage during the data collection process.

Interoperability

To maximize the effectiveness of the FRMIS, standardization of data elements and the development of interface systems to allow seamless data exchange between our HISs were necessary. The FRMIS used Health Level Seven Fast Healthcare Interoperability Resources (HL7FHIR) to enable seamless data exchange and streamline workflows.

Participating Entities

The FRMIS project has obtained the approval and support of hospital management, who have provided strong guarantees in terms of personnel, resources, funding, and working hours required for the implementation of the research plan. Our hospital is an advanced information management hospital with state-of-the-art scientific technologies. The computer center has rich experience in developing information management platforms; they have independently developed and implemented 19 hospital operational management systems. The FRMIS’s development was initiated by the nursing department, with the

assistance of the computer center to fulfill the corresponding requirements.

Budget Planning

The FRMIS development process lasted about 4 months, and the total development cost was approximately 500,000 Renminbi (approximately US $68,300). The subsequent maintenance costs were estimated to be 8% of the total development cost annually. Funding for the FRMIS's development and maintenance was provided by our hospital. The ownership of the FRMIS belongs to Tongji Hospital.

Sustainability

The FRMIS's implementation was carried out through the issuance of relevant policy documents by the nursing department, ensuring its clinical adoption. All risk assessment and incident reports concerning the inpatient falls were conducted through this information system thus far, replacing the previous paper-based forms. Over the past few years of using this system, our hospital's computer center staff has been maintaining and fixing occasional bugs that occur during clinical implementation of this system. The computer center staff also made necessary modifications and improvements to certain details as needed to enhance system functionality, optimize workflows, and adapt to evolving health care practices.

Statistical Analysis

Statistical comparisons were made on the fall incidence rate among inpatients and the reporting rate of high-fall-risk patients before and after FRMIS implementation. Data entry and statistical analysis were performed using SPSS (version 17.0; IBM Corp). The chi-square test was used to compare the differences in the fall incidence rate among inpatients, the rate of high-fall-risk patients, and the implementation rate of preventive fall quality bundle strategy indicators before and after FRMIS implementation. A value of $P<.05$ was considered statistically significant.

Ethics Approval

The study was approved by the institutional review board of Tongji hospital (protocol TJ-IRB20191209).

Implementation (Results)

Coverage

Our hospital is a large academic medical center in central China. In 2016, the hospital had a total of 4000 open beds, 106 nursing wards, and 53 specialized nursing units. The average daily admission rate ranges from 4500 to 5000 patients, with a total of 193,709 admitted patients throughout the year. The cumulative number of bed-days reached 1,756,946, of which 277,365 (15.79%) were for critical patients.

Outcomes

We carried out the process and outcome evaluation with regard to the FRMIS's implementation. The process evaluation indicators include (1) the accuracy rate of the Morse fall risk assessment: number of accurate Morse fall risk assessments / total number of Morse fall risk assessments inspected; (2) implementation rate of fall prevention health education: number of implemented health education check items / number of patients inspected × total number of fall prevention health education check items; (3) implementation rate of shift handoff: number of implemented shift handoff check items / number of patients inspected × total number of fall prevention shift handoff check items; (4) implementation rate of environment safety: number of implemented environment safety check items / the number of patients inspected × the total number of environment safety check items.

The staff of the quality control office in the nursing department reviewed the FRMIS on a daily basis to identify the clinical departments where high-fall-risk patients were distributed across the hospital. For departments with more than 5 high-fall-risk patients and a proportion exceeding 20% of the total patients, we assigned 2 supervisory staff from the quality control team. They used the electronic form "Fall Prevention Bundle Strategy Quality Tracking Form" (see Multimedia Appendix 1) on an iPad to conduct quality inspections on the nursing units for the high-fall-risk patient population, randomly checking the implementation rate of fall prevention bundle strategy indicators (fall risk assessment, fall-related health education, fall-related shift handoff, and environment safety). Before implementing the FRMIS, a total of 1250 patients were randomly sampled for inspection. After implementing FRMIS, a total of 1806 patients were randomly sampled for inspection. Additionally, a comparative analysis was performed on the hospitalization period between February and October 2017 (after FRMIS implementation, the total bed days occupied by inpatients was 1,323,667) and between February and October 2015 (before FRMIS implementation, the total bed days occupied by inpatients was 1,303,094) to evaluate the hospital-wide reporting rate of high-fall-risk cases, incidence rate of patient falls, and severity of fall-related injuries.

The results showed that since the FRMIS's implementation, the inpatient falls rate was significantly lower than that before implementation ($P<.001$), as shown in Table 1. In addition, the percentage of major fall-related injuries was significantly lower than that before implementation, as shown in Table 2. The implementation rate of fall-related process indicators and the reporting rate of high risk of falls were significantly different before and after system implementation ($P<.001$), as shown in Table 3.

Table 1. Comparison of fall-related outcome indicators.

	Before implementation (total bed days=1,303,094), n (%)	After implementation (total bed days=1,323,667), n (%)	Chi-square (*df*)	*P* value
High-fall-risk patients' reports	1036 (0.8)	3007 (2.3)	931.7 (1)	<.001
Fall incident reports	23 (0.02)	11 (0.01)	4.4 (1)	<.001

Table 2. Results of fall-related injuries.

	Cases of fall-related injury, n			
	No injury	Minor	Moderate	Major
Before implementation	15	28	12	2
After implementation	20	13	9	0

Table 3. Comparison of fall-related process indicators.

	Before implementation (n=1250), n (%)	After implementation (n=1806), n (%)	Chi-square (*df*)	*P* value
Fall risk assessment	1056 (84.48)	1709 (95.73)	88 (1)	<.001
Fall-related health education	1107 (88.56)	1769 (97.95)	117.5 (1)	<.001
Fall-related shift handoff	1114 (89.12)	1767 (97.84)	104 (1)	<.001
Environment safety	1127 (90.16)	1796 (99.45)	153 (1)	<.001

Lessons Learned

The FRMIS's development and implementation followed a structured process, starting with needs assessment and culminating in ongoing monitoring and improvement. With this multidisciplinary team and comprehensive approach, we were able to provide a more robust and effective fall risk management system for the entire hospital. The FRMIS addressed the shortcomings of paper-based reporting, such as untimely fall assessments, delayed reporting, information transmission delays, loss of assessment forms, and incomplete tracking information. The FRMIS achieved a holistic fall prevention strategy that spanned from risk assessment to postfall intervention, which brought several benefits to both patients and health care providers. The FRMIS alerted nursing staff about high-risk patients, enabling timely interventions and reducing fall occurrences. It also standardized the reporting process for fall events, allowing for efficient tracking and analysis of incidents.

Discussion

Principal Findings

This study has designed and implemented an FMRIS at the hospital level. The novel system provided a simple, intuitive, and highly operational prevention management model, encompassing fall risk assessment, high fall risk screening, forecasting, and monitoring. It significantly improved the procedural and standardized levels of fall management for hospitalized patients, having prompted nurses to proactively implement fall preventive interventions, conducted timely fall risk assessments, reduced underreporting of high-fall-risk patients, and increased the forecast rate of high-fall-risk patients.

Unlike previous studies that focus on a specific stage of fall management (such as risk identification [19] or fall incident reporting [28]) or patients in a specific department [23], our system catered to the entire process of fall risk management for all inpatients. The FRMIS showed promise in enhancing patient safety, reducing fall incidents, and improving overall care quality.

To facilitate the successful implementation of the FRMIS in clinical practice, we first developed the Standardized Management Guidelines for Preventing Inpatient Falls at the hospital-wide level. This policy document comprehensively revises and improves clinical fall prevention efforts, which include patient fall risk assessment, health education, fall preventive interventions, fall management workflow, fall incident reporting, and system record-keeping. The policy document was distributed in hard copy by the nursing department to all departments and also uploaded electronically on the hospital's Office Automation platform. It mandated each clinical department to conduct fall prevention training based on the guidelines, requiring all nurses' participation and proficiency. This document served as a supporting tool, providing nurses with guidance on how to use the FRMIS effectively in their clinical practice to prevent inpatient falls.

In addition, we conducted standardized nurse training through a web-based platform. Three main implementation strategies were used. First, we conducted diverse forms of training, including ward-, department-, and hospital-level fall prevention training, as well as case-based warning education, bedside simulation assessment, experience sharing sessions, and special lectures, to comprehensively implement the content of the Standardized Management Guidelines for Preventing Falls. Second, we performed objective evaluation. We incorporated simulated case examinations for patient fall prevention into the clinical skills evaluation of nurses, head nurse position evaluation, and their performance appraisal to comprehensively assess the level of knowledge of fall management guidelines and the emergency handling capabilities for patient fall incidents. Third, we achieved full participation among all nurses. The training rate and assessment results of nurses in the wards were included in the performance management projects of ward head nurses, achieving the participation of all nurses and comprehensive evaluation of standardized fall prevention training. Based on the strategies mentioned above, the FRMIS's implementation in clinical practice has been relatively successful.

Limitations

This study still has certain limitations that should be acknowledged. First, the FRMIS was specifically designed and implemented by our hospital's computer center. It is currently applicable to 3 different hospital campuses within our institution

but has not been widely disseminated to other hospitals or integrated with diverse HISs. Therefore, its applicability and effectiveness in different hospital contexts remains uncertain. Second, the FRMIS heavily relied on the voluntary reporting by clinical nurses. The accuracy of these fall risk reports needed to be individually verified by staff members in the quality control office of the nursing department. This process is currently manual and lacks automation, which may introduce delays and potential inconsistencies. In the future, further improvements could be made by integrating artificial intelligence (AI) technologies. By automatically extracting fall risk factors from patients' electronic medical records, the system could achieve automated risk stratification and reduce dependence on manual reporting.

Despite these limitations, it is important to note that this study represents a significant step toward enhancing inpatient fall risk management through the FRMIS implementation. Future research and development efforts could focus on expanding the system's applicability to other hospitals, integrating AI capabilities for automated risk assessment, and improving data accuracy and automation processes. These advancements would contribute to more comprehensive and intelligent fall risk management practices for inpatients.

Conclusions

The design and implementation of an FRMIS significantly contributed to the prevention and management of falls among inpatients. The FRMIS enhanced patient safety through IT, providing comprehensive support for fall prevention and ensuring efficient management of fall events in health care settings.

Acknowledgments

We sincerely thank the nursing management and the participating nurses of the Tongji hospital for their support and participation in this study. This study was partly funded by the Huazhong University of Science and Technology Independent Innovation Fund (2013YQ008, 2018KFYYXJJ016), Chinese Nursing Association Research Project (ZHKY202204), and China Nursing Management Research Fund (CNM-2020-03).

Data Availability

The data sets used or analyzed during this study available from the corresponding author on reasonable request.

Authors' Contributions

WY designed the study. JMY and HM collected the data. HM and DMJ analyzed the data. JMY wrote the original draft of the manuscript. WY and HM reviewed and edited the manuscript. WY applied for funding. All authors have read and agreed to the version of the manuscript intended for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The presentation of the fall risk management information system.
[DOCX File , 766 KB - medinform_v12i1e46501_app1.docx]

References

1. Oner B, Zengul FD, Oner N, Ivankova NV, Karadag A, Patrician PA. Nursing-sensitive indicators for nursing care: A systematic review (1997-2017). Nurs Open 2021 May;8(3):1005-1022 [FREE Full text] [doi: 10.1002/nop2.654] [Medline: 34482649]
2. James S, Lucchesi L, Bisignano C, Castle C, Dingels Z, Fox J, et al. The global burden of falls: global, regional and national estimates of morbidity and mortality from the Global Burden of Disease Study 2017. Inj Prev 2020 Oct;26(Supp 1):i3-i11 [FREE Full text] [doi: 10.1136/injuryprev-2019-043286] [Medline: 31941758]
3. Burns ER, Stevens JA, Lee R. The direct costs of fatal and non-fatal falls among older adults - United States. J Safety Res 2016 Sep;58:99-103 [FREE Full text] [doi: 10.1016/j.jsr.2016.05.001] [Medline: 27620939]
4. Eslami Andargoli A, Scheepers H, Rajendran D, Sohal A. Health information systems evaluation frameworks: a systematic review. Int J Med Inform 2017 Jan;97:195-209. [doi: 10.1016/j.ijmedinf.2016.10.008] [Medline: 27919378]
5. Gaspar AGM, Lapão LV. eHealth for addressing balance disorders in the elderly: systematic review. J Med Internet Res 2021 Apr 28;23(4):e22215 [FREE Full text] [doi: 10.2196/22215] [Medline: 33908890]
6. Field M, Fong K, Shade C. Use of electronic visibility boards to improve patient care quality, safety, and flow on inpatient pediatric acute care units. J Pediatr Nurs 2018 Jul;41:69-76. [doi: 10.1016/j.pedn.2018.01.015] [Medline: 29395791]

7. Woodward M, De Pennington N, Grandidge C, McCulloch P, Morgan L. Development and evaluation of an electronic hospital referral system: a human factors approach. Ergonomics 2020 Jun;63(6):710-723. [doi: 10.1080/00140139.2020.1748232] [Medline: 32220218]
8. Chow CB, Leung M, Lai A, Chow YH, Chung J, Tong KM, et al. Development of an electronic emergency department-based geo-information injury surveillance system in Hong Kong. Injury 2012 Jun;43(6):739-748. [doi: 10.1016/j.injury.2011.08.008] [Medline: 21924722]
9. Carrillo I, Mira JJ, Vicente MA, Fernandez C, Guilabert M, Ferrús L, et al. Design and testing of BACRA, a web-based tool for middle managers at health care facilities to lead the search for solutions to patient safety incidents. J Med Internet Res 2016 Sep 27;18(9):e257 [FREE Full text] [doi: 10.2196/jmir.5942] [Medline: 27678308]
10. Balaguera HU, Wise D, Ng CY, Tso H, Chiang W, Hutchinson AM, et al. Using a medical intranet of things system to prevent bed falls in an acute care hospital: a pilot study. J Med Internet Res 2017 May 04;19(5):e150 [FREE Full text] [doi: 10.2196/jmir.7131] [Medline: 28473306]
11. Dalal AK, Fuller T, Garabedian P, Ergai A, Balint C, Bates DW, et al. Systems engineering and human factors support of a system of novel EHR-integrated tools to prevent harm in the hospital. J Am Med Inform Assoc 2019 Jun 01;26(6):553-560 [FREE Full text] [doi: 10.1093/jamia/ocz002] [Medline: 30903660]
12. Stockwell-Smith G, Adeleye A, Chaboyer W, Cooke M, Phelan M, Todd J, et al. Interventions to prevent in-hospital falls in older people with cognitive impairment for further research: a mixed studies review. J Clin Nurs 2020 Sep;29(17-18):3445-3460. [doi: 10.1111/jocn.15383] [Medline: 32578913]
13. Tricco AC, Thomas SM, Veroniki AA, Hamid JS, Cogo E, Strifler L, et al. Quality improvement strategies to prevent falls in older adults: a systematic review and network meta-analysis. Age Ageing 2019 May 01;48(3):337-346 [FREE Full text] [doi: 10.1093/ageing/afy219] [Medline: 30721919]
14. LeLaurin JH, Shorr RI. Preventing falls in hospitalized patients: state of the science. Clin Geriatr Med 2019 May;35(2):273-283 [FREE Full text] [doi: 10.1016/j.cger.2019.01.007] [Medline: 30929888]
15. Cameron ID, Dyer SM, Panagoda CE, Murray GR, Hill KD, Cumming RG, et al. Interventions for preventing falls in older people in care facilities and hospitals. Cochrane Database Syst Rev 2018 Sep 07;9(9):CD005465 [FREE Full text] [doi: 10.1002/14651858.CD005465.pub4] [Medline: 30191554]
16. Ambrens M, Tiedemann A, Delbaere K, Alley S, Vandelanotte C. The effect of eHealth-based falls prevention programmes on balance in people aged 65 years and over living in the community: protocol for a systematic review of randomised controlled trials. BMJ Open 2020 Jan 15;10(1):e031200 [FREE Full text] [doi: 10.1136/bmjopen-2019-031200] [Medline: 31948985]
17. Turner K, Staggs V, Potter C, Cramer E, Shorr R, Mion LC. Fall prevention implementation strategies in use at 60 United States hospitals: a descriptive study. BMJ Qual Saf 2020 Dec;29(12):1000-1007 [FREE Full text] [doi: 10.1136/bmjqs-2019-010642] [Medline: 32188712]
18. Morgan L, Flynn L, Robertson E, New S, Forde-Johnston C, McCulloch P. Intentional Rounding: a staff-led quality improvement intervention in the prevention of patient falls. J Clin Nurs 2017 Jan;26(1-2):115-124. [doi: 10.1111/jocn.13401] [Medline: 27219073]
19. Giles LC, Whitehead CH, Jeffers L, McErlean B, Thompson D, Crotty M. Falls in hospitalized patients: can nursing information systems data predict falls? Comput Inform Nurs 2006;24(3):167-172. [doi: 10.1097/00024665-200605000-00014] [Medline: 16707948]
20. Mei YY, Marquard J, Jacelon C, DeFeo AL. Designing and evaluating an electronic patient falls reporting system: perspectives for the implementation of health information technology in long-term residential care facilities. Int J Med Inform 2013 Nov;82(11):e294-e306. [doi: 10.1016/j.ijmedinf.2011.03.008] [Medline: 21482183]
21. Katsulis Z, Ergai A, Leung WY, Schenkel L, Rai A, Adelman J, et al. Iterative user centered design for development of a patient-centered fall prevention toolkit. Appl Ergon 2016 Sep;56:117-126. [doi: 10.1016/j.apergo.2016.03.011] [Medline: 27184319]
22. Dykes PC, Duckworth M, Cunningham S, Dubois S, Driscoll M, Feliciano Z, et al. Pilot Testing Fall TIPS (Tailoring Interventions for Patient Safety): a Patient-Centered Fall Prevention Toolkit. Jt Comm J Qual Patient Saf 2017 Aug;43(8):403-413. [doi: 10.1016/j.jcjq.2017.05.002] [Medline: 28738986]
23. Jacobsohn GC, Leaf M, Liao F, Maru AP, Engstrom CJ, Salwei ME, et al. Collaborative design and implementation of a clinical decision support system for automated fall-risk identification and referrals in emergency departments. Healthc (Amst) 2022 Mar;10(1):100598 [FREE Full text] [doi: 10.1016/j.hjdsi.2021.100598] [Medline: 34923354]
24. Mlaver E, Schnipper JL, Boxer RB, Breuer DJ, Gershanik EF, Dykes PC, et al. User-centered collaborative design and development of an inpatient safety dashboard. Jt Comm J Qual Patient Saf 2017 Dec;43(12):676-685. [doi: 10.1016/j.jcjq.2017.05.010] [Medline: 29173289]
25. 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]
26. Morse J. Preventing Patient Falls. Thousand Oaks, CA: Sage Publications; 1997.

27. Schwendimann R, De Geest S, Milisen K. Evaluation of the Morse Fall Scale in hospitalised patients. Age Ageing 2006 May;35(3):311-313. [doi: 10.1093/ageing/afj066] [Medline: 16527829]
28. Gardner LA, Bray PJ, Finley E, Sterner C, Ignudo TL, Stauffer CL, et al. Standardizing falls reporting: using data from adverse event reporting to drive quality improvement. J Patient Saf 2019 Jun;15(2):135-142. [doi: 10.1097/PTS.0000000000000204] [Medline: 26332598]

Abbreviations

AI: artificial intelligence
FRMIS: fall risk management information system
HIS: hospital information system
HL7FHIR: Health Level Seven Fast Healthcare Interoperability Resources
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
TIPS: Tailoring Interventions for Patient Safety

Edited by C Perrin; submitted 14.02.23; peer-reviewed by M Binandeh, R Hu, A Krupp; comments to author 06.04.23; revised version received 15.08.23; accepted 29.11.23; published 02.01.24.

Please cite as:
Wang Y, Jiang M, He M, Du M
Design and Implementation of an Inpatient Fall Risk Management Information System
JMIR Med Inform 2024;12:e46501
URL: https://medinform.jmir.org/2024/1/e46501
doi:10.2196/46501
PMID:38165733

Implementation Report

A Nationwide Chronic Disease Management Solution via Clinical Decision Support Services: Software Development and Real-Life Implementation Report

Mustafa Mahir Ulgu[1], MD, PhD; Gokce Banu Laleci Erturkmen[2], PhD; Mustafa Yuksel[2], PhD; Tuncay Namli[2], PhD; Şenan Postacı[2], MSc; Mert Gencturk[2], PhD; Yildiray Kabak[2], PhD; A Anil Sinaci[2], PhD; Suat Gonul[2], PhD; Asuman Dogac[2], PhD; Zübeyde Özkan Altunay[1], MD; Banu Ekinci[1], MD; Sahin Aydin[1], MSc; Suayip Birinci[1], MD

[1]Ministry of Health Turkey, Ankara, Turkey

[2]Software Research Development and Consultancy Corporation, Ankara, Turkey

Corresponding Author:
Gokce Banu Laleci Erturkmen, PhD
Software Research Development and Consultancy Corporation
Orta Dogu Teknik Universitesi Teknokent Silikon Blok Kat 1 No 16
Ankara, 06800
Turkey
Phone: 90 3122101763
Email: gokce@srdc.com.tr

Abstract

Background: The increasing population of older adults has led to a rise in the demand for health care services, with chronic diseases being a major burden. Person-centered integrated care is required to address these challenges; hence, the Turkish Ministry of Health has initiated strategies to implement an integrated health care model for chronic disease management. We aim to present the design, development, nationwide implementation, and initial performance results of the national Disease Management Platform (DMP).

Objective: This paper's objective is to present the design decisions taken and technical solutions provided to ensure successful nationwide implementation by addressing several challenges, including interoperability with existing IT systems, integration with clinical workflow, enabling transition of care, ease of use by health care professionals, scalability, high performance, and adaptability.

Methods: The DMP is implemented as an integrated care solution that heavily uses clinical decision support services to coordinate effective screening and management of chronic diseases in adherence to evidence-based clinical guidelines and, hence, to increase the quality of health care delivery. The DMP is designed and implemented to be easily integrated with the existing regional and national health IT systems via conformance to international health IT standards, such as Health Level Seven Fast Healthcare Interoperability Resources. A repeatable cocreation strategy has been used to design and develop new disease modules to ensure extensibility while ensuring ease of use and seamless integration into the regular clinical workflow during patient encounters. The DMP is horizontally scalable in case of high load to ensure high performance.

Results: As of September 2023, the DMP has been used by 25,568 health professionals to perform 73,715,269 encounters for 16,058,904 unique citizens. It has been used to screen and monitor chronic diseases such as obesity, cardiovascular risk, diabetes, and hypertension, resulting in the diagnosis of 3,545,573 patients with obesity, 534,423 patients with high cardiovascular risk, 490,346 patients with diabetes, and 144,768 patients with hypertension.

Conclusions: It has been demonstrated that the platform can scale horizontally and efficiently provides services to thousands of family medicine practitioners without performance problems. The system seamlessly interoperates with existing health IT solutions and runs as a part of the clinical workflow of physicians at the point of care. By automatically accessing and processing patient data from various sources to provide personalized care plan guidance, it maximizes the effect of evidence-based decision support services by seamless integration with point-of-care electronic health record systems. As the system is built on international code systems and standards, adaptation and deployment to additional regional and national settings become easily possible. The nationwide DMP as an integrated care solution has been operational since January 2020, coordinating effective screening and management of chronic diseases in adherence to evidence-based clinical guidelines.

(JMIR Med Inform 2024;12:e49986) doi:10.2196/49986

KEYWORDS

chronic disease management; clinical decision support services; integrated care; interoperability; evidence-based medicine; medicine; disease management; management; implementation; decision support; clinical decision; support; chronic disease; physician-centered; risk assessment; tracking; diagnosis

Introduction

As in the rest of the world, the aging population is increasing rapidly in Turkey. A recent TurkStat report predicts that by 2030, the older adult population will be 12.9%, rising to 22.6% in 2060 and 25.6% in 2080 [1]. Noncommunicable diseases are the leading cause of death and disability in Turkey, posing a significant burden [2]. The elevated health costs for older adults strain Turkey's health care system. To address this, the Turkish Ministry of Health (MOH) has implemented a national strategy emphasizing multidisciplinary teams, led by family physicians. The goal is to enhance early detection and manage complications of noncommunicable diseases through systematic screening programs under the national Disease Management Platform (DMP) project launched in late 2018.

The growing use of digital health solutions such as electronic health records (EHRs) presents an opportunity to enhance chronic disease management. Clinical decision support services (CDSSs) can assist in making patient-centered and evidence-based decisions [3,4]. Digital tools and systems that collect and use patient information to provide decision support for health care professionals (HCPs), including patient-specific assessments and recommendations, can promote adherence to national guidelines, ultimately resulting in enhanced quality of care [5-9]. Research demonstrated that computerized decision support tailored to the patient successfully improved decision-making [10,11]. Such tools enhanced the decision-making abilities of HCPs in various domains, including effective prescription decisions [12,13], adherence to guidelines for cardiac rehabilitation [14], management of hypertension and diabetes [15-21], cancer screening [22,23], and computerized order decisions [24,25].

Building on these results, the national DMP is designed as an integrated care platform for chronic disease management in Turkey in a family physician–centered manner. It aims to effectively implement clinical treatment protocols, ensuring easy adherence with decision support services. These services focus on early diagnosis, followed by structured treatment recommendations during routine follow-ups. The DMP enhances standardization of care, improving health care efficiency and quality. It also facilitates seamless transitions between primary care and specialist services, reducing costs, minimizing risks, eliminating redundant tests, and easing the burden on patients.

To ensure successful implementation of a DMP aimed at achieving these strategic objectives, several technical challenges need to be addressed. Our design decisions consider the crucial factor of integrating CDSSs seamlessly into clinicians' daily workflow [26,27]. Despite the potential of CDSSs for evidence-based medicine, significant effort is needed to realize these benefits [28]. The DMP must smoothly integrate with physicians' workflows, necessitating interoperability with existing health IT systems. CDSS guidance should be user-friendly, ensuring a natural flow for clinical protocol implementation. With a target audience of over 26,000 practitioners in Turkey, serving a population of over 85 million, the platform must ensure high performance and scalability. It should easily expand to address additional diseases within a reasonable timeframe and prioritize reusability and compliance with international health IT standards for versatile deployment.

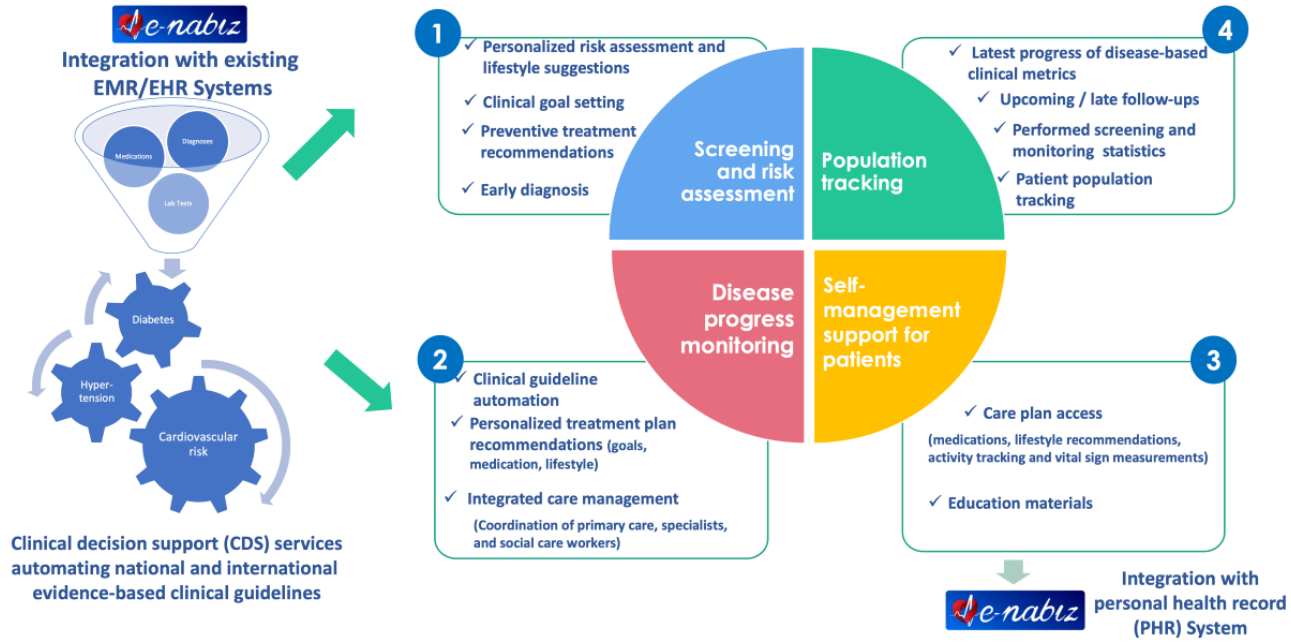
This paper outlines the design, development, nationwide implementation, and initial performance results of the national DMP in Turkey. The DMP can be categorized as a "2.3-Healthcare Provider Decision Support System" in terms of World Health Organization "Classification of digital health interventions" [29]. This implementation report will focus on the results of the deployment and implementation of the DMP in Turkey serving to more than 26,000 family medicine practitioners (FMPs) in the country. The objective is to share our experiences in building the DMP, as an implementation report in line with *iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations* [30]. We detail the design decisions and technical solutions aimed at ensuring interoperability with existing IT systems, integration with clinical workflow, enabling smooth transition of care, user-friendliness for HCPs, scalability, and adaptability in the *Methods* section. The *Implementation (Results)* section presents the outcomes of the nationwide implementation (number of users, number of screening and monitoring encounters, number of patients covered via these encounters, number of patients diagnosed as a result of screening encounters, and treatment goal achievements [such as blood pressure targets, hemoglobin A_{1c} {HbA_{1c}}, and cholesterol targets]), demonstrating how these objectives were achieved. Additionally, we outline current limitations and identify areas for future work to further enhance the clinical impact.

Methods

Overall System Architecture and Design Decisions

The DMP has been designed and implemented to enable the following 4 high-level features as summarized in Figure 1:

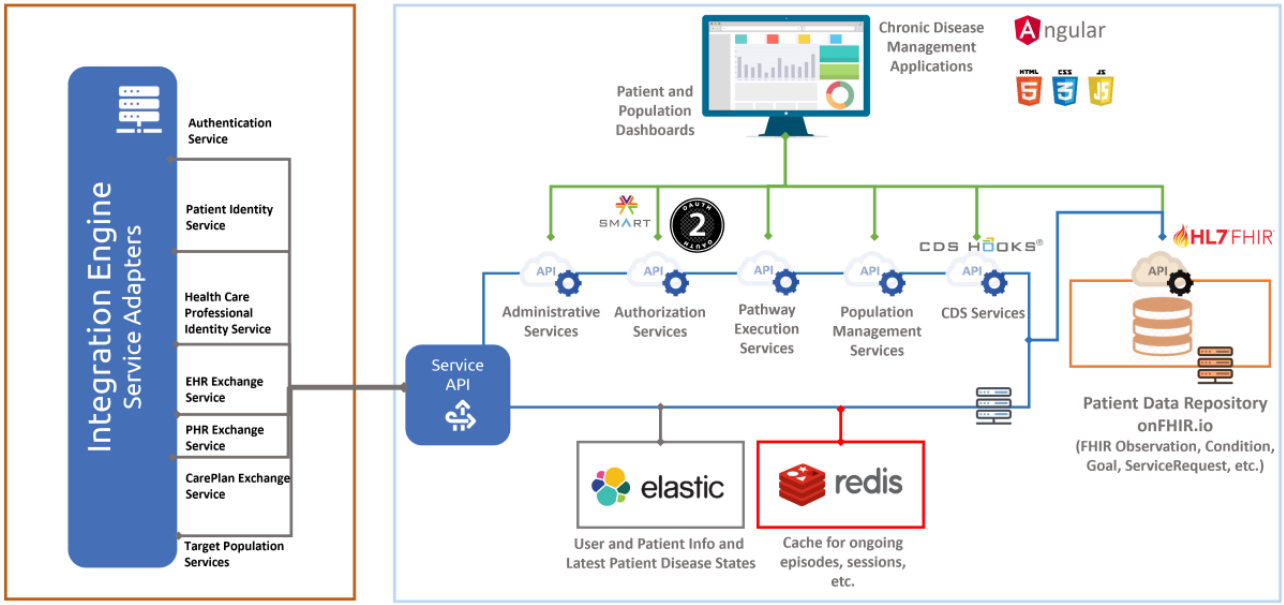
Figure 1. Overall aims of the disease management platform architecture. EHR: electronic health record; EMR: electronic medical record.



- Screening and risk assessment for healthy population: a web-based platform for FMPs facilitates screening for the healthy population. For instance, diabetes screening is required every 3 years for citizens aged over 40 years without a diabetes diagnosis. The full eligibility criteria for both screening and monitoring are presented in Multimedia Appendix 1. The system offers personalized risk assessments, early diagnosis, individualized goals, preventive treatment, and lifestyle suggestions aligned with national care pathways. Diagnosed patients enter the disease progress monitoring program, whereas undiagnosed individuals receive intensified screening based on risk and lifestyle recommendations.
- Disease progress monitoring: for diagnosed patients, the platform facilitates creating and updating personalized care plans during regular follow-up encounters, aligning with evidence-based national care pathways. It assesses laboratory results, conducts risk assessments, recommends personalized treatment goals and medications, suggests follow-up appointments, and refers to specialists when necessary for consultations and complication management. Patients in the monitoring program are categorized based on their control of clinical parameters, symptoms, and goal achievement status, guiding decisions on follow-up frequency, secondary care referrals, and medication plan updates.
- Self-management support for patients: a care plan with instructions for FMPs, specialists, and patients is shared with Turkey's e-Nabiz platform, the national EHR and personal health record (PHR) system. Patients can then access details about care plan activities, including medications, educational materials, self-measurement activities, and lifestyle recommendations.
- Population tracking: each FMP manages 2000 to 4000 patients based on their region's population. The population tracking module allows them to filter and manage patients for upcoming or overdue screening and monitoring encounters, access statistics on the screened population, send SMS invitations to patients, and monitor goal achievement for clinical parameters such as fasting plasma glucose, HbA_{1c}, and blood pressure.

The overall system architecture of the DMP is depicted in Figure 2.

Figure 2. High-level system architecture of the disease management platform. API: application programming interface; CDS: clinical decision support; EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven; PHR: personal health record.



Seamless Integration and Interoperability With Existing Systems

The DMP is designed and implemented for seamless integration with existing regional and national health IT systems. To achieve this, we have designed the core data model and data processing architecture of the DMP based on Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) Release 4 [31]. FHIR has gained widespread adoption in the health care industry [32-36] and endorsed by country-wide implementations in the United States [34], United Kingdom [37], and Germany [38].

The DMP core data model conforms to HL7 FHIR Release 4 to encompass basic EHR components as well as resources for representing a patient's care plan. An open-source HL7 FHIR Repository, namely onFHIR.io [39], serves as the main component of the data management layer (Figure 2). onFHIR.io uses MongoDB as a database and provides real-time data subscription with the help of Apache Kafka. The DMP web application directly accesses patient and care plan data through RESTful interfaces provided by onFHIR.io, enabling fine-grained access control over all FHIR resources in compliance with the SMART on FHIR authorization guidelines and scopes [40].

In Turkey, the MOH operates e-Nabiz, a central national EHR infrastructure [41]. This system collects patient records as encounter summaries from nationwide health care providers, with patients also inputting vital signs and activity data. e-Nabiz codes data using international and national medical terminology, such as *International Classification of Diseases, Tenth Revision* (ICD-10). It is a document-based repository accessed through a Representational State Transfer application programming interface [42], and interoperability adapters in the DMP project (EHR exchange and PHR exchange services in Figure 2) communicate with it to retrieve patient data. These adapters transform proprietary XML formats to HL7 FHIR-based data models and store them in the Patient Data Repository. This transformation includes both structural and semantic mapping, incorporating a strategy of incremental synchronization. On initial DMP access, the patient's longitudinal EHR is mapped to FHIR, and subsequent encounters retrieve and transform only new, unsynchronized data.

To secure patient data access, clinicians authenticate to the DMP through the MOH's central authentication and authorization services using the OpenID Connect protocol. The DMP uses a role-based access control mechanism, catering to different disease management roles. Before data access and synchronization, a check ensures that the user has the required access rights via the MOH's central authentication service. If authorized, the DMP generates a patient-specific JavaScript Object Notation Web Token with corresponding permissions, serving as an OAuth2.0 bearer token for all interactions within the DMP.

In the DMP, FMPs perform screening and monitoring encounters based on predefined eligibility criteria. For instance, hypertension monitoring is required every 3 months for patients with a hypertension diagnosis and on antihypertensive medications. These criteria are executed in the e-Nabiz data warehouse, and both DMP and family medicine information systems retrieve target population lists through target population services (Figure 2). FMPs can easily identify if a visiting patient is on the screening or monitoring list via family medicine information systems, initiating a DMP encounter directly with a single sign-on integration.

The care plan created with the help of the DMP is stored as an HL7 FHIR *CarePlan* resource in the Patient Data Repository. It is shared with the e-Nabiz system via the Care Plan Exchange Service (Figure 2), enabling it to be accessible to the patient via e-Nabiz interfaces.

The DMP uses Elasticsearch technology for storing user information, basic patient attributes, and their current screening and monitoring statuses for each disease. Elasticsearch also serves as a system log repository. We have developed a Kibana

interface for monitoring system performance and geographical statistics. Redis is used as a caching system to temporally store information about ongoing encounters and user authorization access tokens.

Automation of National Care Pathways as a Clinical Workflow for FMPs

The interfaces of the DMP have been designed with ease of use in mind to allow for seamless integration into the regular clinical workflow. It is implemented as a cocreation activity with the involvement of system analysts, software engineers, and a clinical reference group set up by the MOH Department of Chronic Diseases and Elderly Health including multidisciplinary HCPs.

The national evidence-based care pathways have been collaboratively analyzed, leading to the identification of common steps, such as physical examination, medical history review, risk assessment, medication review, lab results review, diagnosis, clinical goal setting, pharmacological treatment planning, and nonpharmacological treatment planning. Each care pathway is designed modularly within the DMP as a series of pages corresponding to these common steps. These are organized as a flow of pages that is followed automatically based on patient parameters.

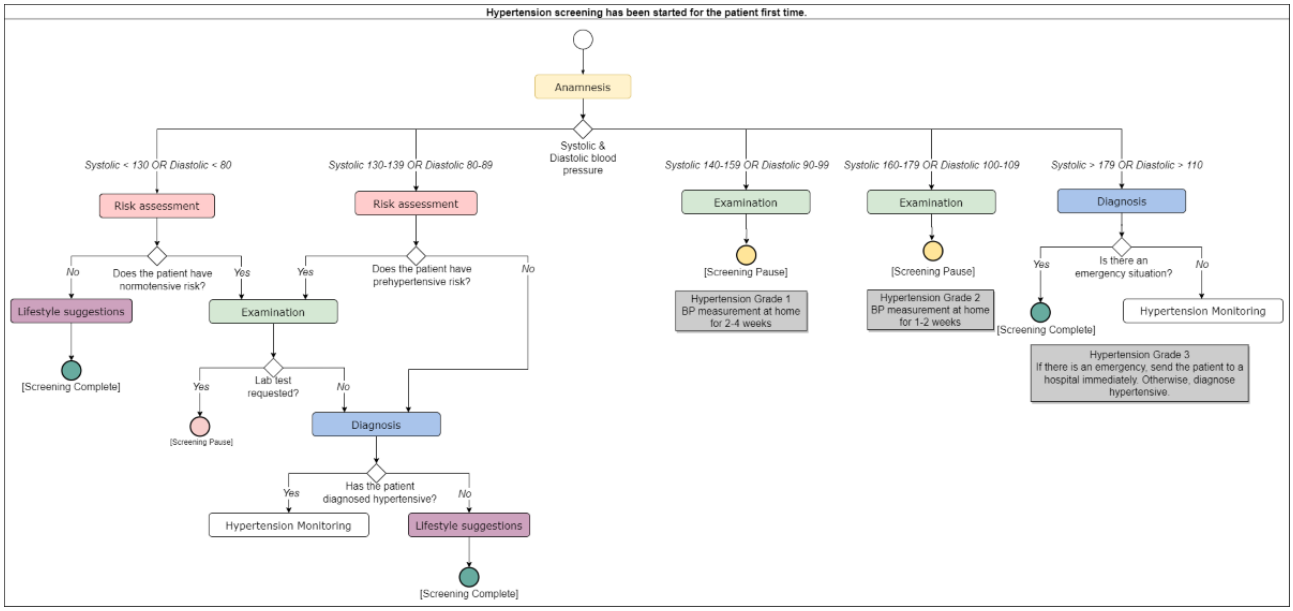
Each page is meticulously designed, specifying patient parameters for assessment. Most data come from the national EHR system, enabling clinicians to review prefilled pages with the latest parameters and make adjustments as needed. Validity periods for each parameter are identified, emphasizing recency, and they are enforced by the system and reminded to FMPs. Additionally, scaled assessments (eg, Mini Nutritional Assessment), risk assessments (eg, cardiovascular risk), and associated algorithms (eg, SCORE-Turkey) are also identified. Business rules within the pages are designed for personalized suggestions aligned with evidence-based care pathways.

All of these are thoroughly documented after discussions in cocreation workshops. Mock-up screens are designed, and flow diagrams are created to identify transition criteria between pages. These materials undergo further review and finalization in subsequent cocreation workshops. As an illustrative example, Figure 3 depicts a sample flow for hypertension screening.

After cocreation, each step's design becomes a web-based interface in the DMP application, developed with the Angular framework. A "Pathway Execution Service" state manager automates the flow diagram for disease screening or monitoring, adapting to patient parameters. This allows FMPs to use a wizard-like interface for encounters, facilitating adherence to national clinical pathways.

Transitions between disease modules are also modeled and implemented. For example, in hypertension screening, if a patient's fasting plasma glucose exceeds 110 mg/dL, the system prompts FMPs to consider a diabetes screening if not already monitored for diabetes. In response, the patient's diabetes screening schedule is automatically updated.

Figure 3. Hypertension screening flow. BP: blood pressure.



CDSS Implementation

CDSSs are a core component of DMP to enable patient-tailored recommendations. On the basis of the documented business rules from the design phase, we have designed CDSSs as automated processes. These processes link patient-specific data with evidence-based knowledge from national care pathways. We can categorize the CDSS implemented based on their functionality as follows:

- Risk assessment via scored algorithms (eg, SCORE-Turkey): FMPs are provided with explanatory guidance about scoring, referencing validated scoring assessment algorithms (see Figure 4).
- Diagnosis recommendations based on the patient's current condition and risk assessment: in screening operations, the CDSS recommends diagnoses to FMPs using predefined ICD-10 codes.

- Guidance for lab test ordering and interpretation: a personalized list of required lab tests is determined based on the patient's disease state, risks, and other comorbidities. The CDSS also provides notifications for when these lab tests should be renewed on expiration.
- Diagnosis and referral suggestions are recommended based on patient parameters such as lab results. For example, referral to a nephrologist is recommended when the estimated glomerular filtration rate result is below 60 mL/min/1.73 m^2.
- Treatment goals (eg, low-density lipoprotein cholesterol) are recommended based on the patient's risk, disease stage, and comorbidities. In Figure 5, an example screen for goal planning is presented. The physician can always manually update these targets based on their assessments.
- Medication suggestions are recommended for treatment planning based on disease stage, response to previous medications, existing medications, and comorbidities. Certain medications are marked as contraindications based on the existing comorbidities of the patient.
- Referral suggestions for preventive consultation visits are recommended, especially for complication management. For instance, a yearly retinopathy check with an ophthalmologist is advised during diabetes monitoring encounters.
- Follow-up visits are recommended based on the current status of the patient. For instance, screening in each 2 years is suggested for patients with low cardiovascular disease risk, whereas once a year screening is suggested for high-risk patients.
- Automated care pathway transitions for patients with multiple morbidities are personalized based on specific disease criteria. For instance, if a patient aged over 40 years has not had their cardiovascular risk score calculated, the DMP guides FMPs to continue with the cardiovascular risk module during hypertension or diabetes monitoring.

In the DMP, all CDSS implementations adhere to the CDS Hooks specifications [43]. As a standard published by HL7, it provides an API specification for CDS calls. Both input parameters and output suggestions are defined in reference to HL7 FHIR resources, facilitating plug-and-play interoperability with platforms that support HL7 FHIR. The CDS Hooks–compliant approach allows easy expansion with CDSSs created by external entities and to simplify deployment in different settings already adhering to HL7 FHIR.

Figure 4. An example screenshot from the Cardiovascular Risk Screening Module presenting individualized risk calculation. (The system is implemented in a multilingual manner supporting Turkish and English by default.) CVD: cardiovascular disease.

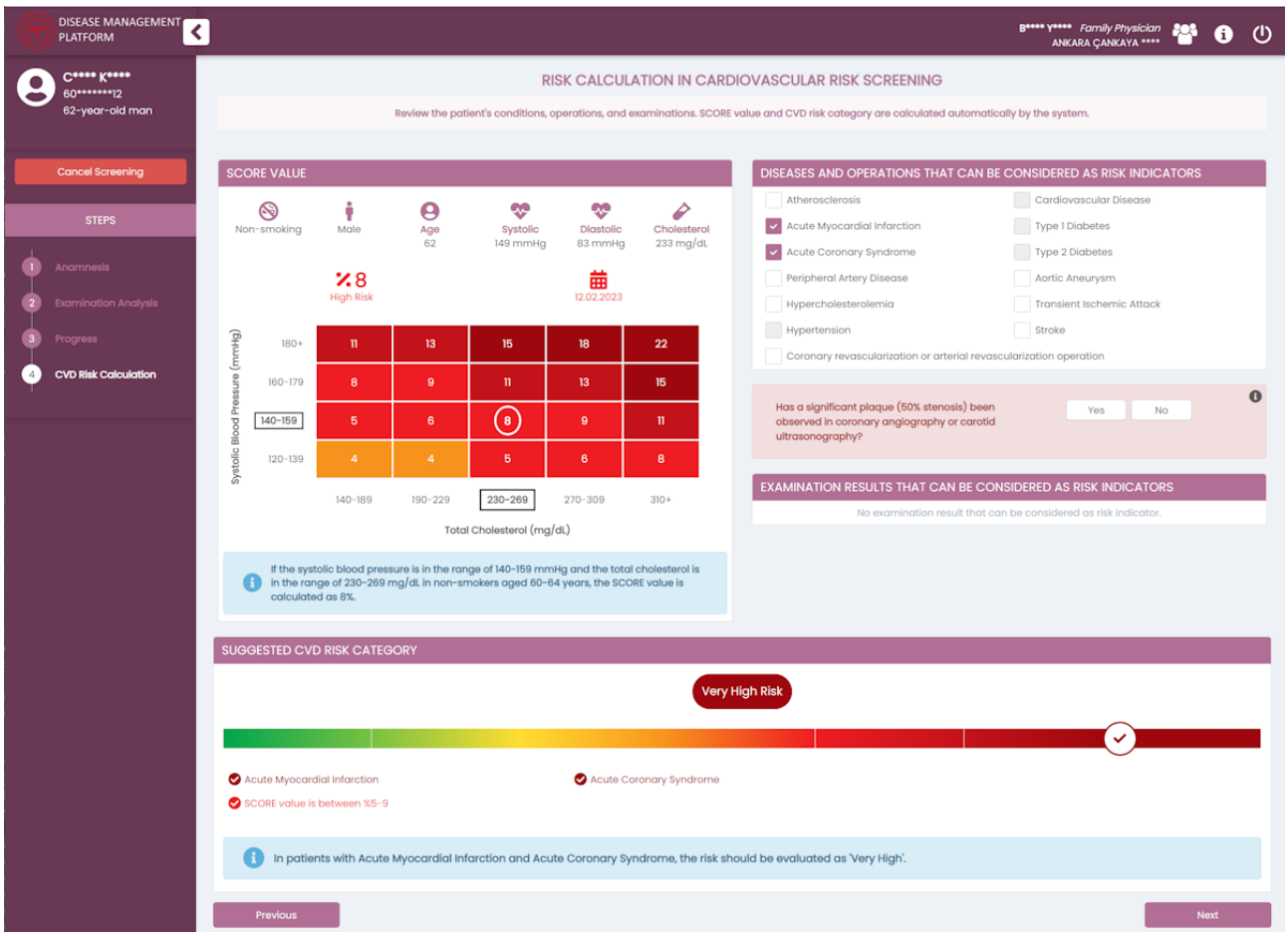
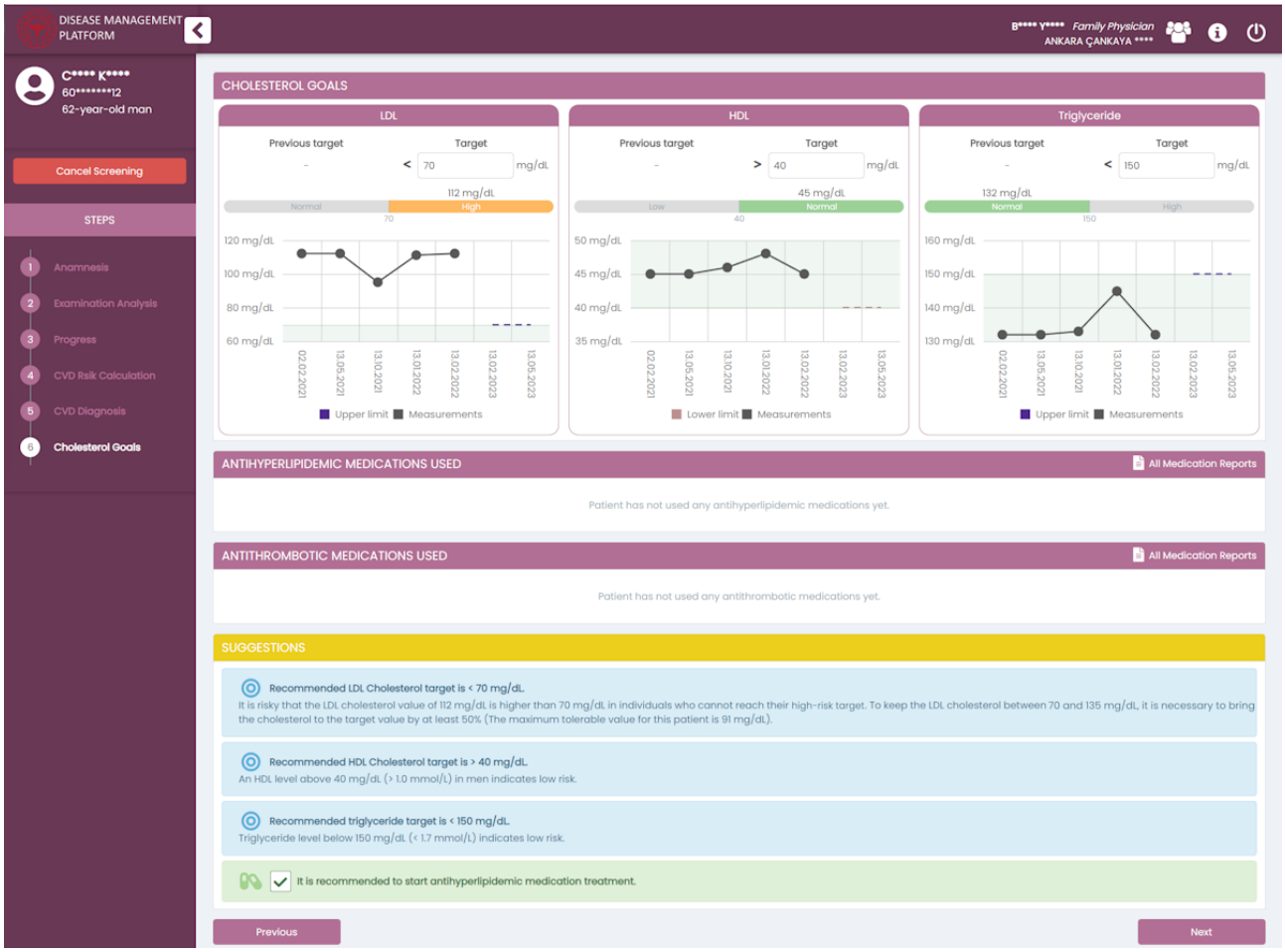


Figure 5. An example screenshot from the disease management platform presenting personalized lipid goals for the patient. HDL: high-density lipoprotein; LDL: low-density lipoprotein.



Ensuring Performance of the System

The DMP is designed for high horizontal scalability, using 2 servers for the web application and 3 for the Patient Data Repository, forming an onFHIR.io server cluster. Nginx acts as both a reverse proxy and a load balancer to distribute traffic across these backend servers. onFHIR.io servers connect to a horizontally scalable MongoDB cluster for data distribution and replication. Elasticsearch log and data store operate on a cluster hosted on 6 servers.

Testing, Piloting, and Deployment of Disease Modules

The system is developed by SRDC on behalf of the Turkish MOH with the support of Türksat and Innova. The final product is owned by the Turkish MOH. The initial version of the DMP, including modules for screening and monitoring of type 2 diabetes, hypertension, and cardiovascular risk management, was extensively tested by the clinical reference group. It underwent a 3-month pilot phase in 4 cities in late 2019. The pilot phase involved 14,351 encounters conducted by 219 FMPs for 5521 patients. Two more modules for obesity screening and monitoring, as well as older adult monitoring, were added to the system during this period. After the feedback is addressed and the system is retested, the system has been operationalized in whole Turkey by January 2020. On June 30, 2021, the MOH has published a directive incentivizing FMPs to conduct screening and monitoring for diabetes, hypertension, cardiovascular disease risk management, obesity, and older adult monitoring via the DMP. The system with incentivization calculations has been operational in whole Turkey since July 1, 2021.

Beyond existing modules, the system now includes monitoring modules for coronary artery disease, chronic kidney disease, stroke, chronic obstructive pulmonary disease, and asthma. The cocreation process, covering requirement analysis, mock-up design, implementation, and testing, took 3 months for each module, showcasing the process's repeatability for swift module additions. These new disease modules are not yet public in the operational system.

Implementation (Results)

The system is being used extensively throughout the whole country. As of September 18, 2023, a total of 73,715,269 screening and monitoring encounters have been performed by 25,568 users (24,627 FMPs and 941 FMP nurses) for 16,058,904 unique citizens. Among these citizens, 56.2% (n=9,025,104) are female and 43.8% (n=7,033,800) are male. The average number of DMP encounters per patient is 4.59. The distribution of encounters per DMP module and the breakdown between screening and monitoring is provided in Table 1.

In Turkey, there are 26,600 FMP units, with each unit using 1 FMP at a time. As of September 18, 2023, FMPs working at

26,210 (98.5%) unique FMP units have logged into the DMP at least once, and 22,982 (86.4%) FMP units have performed at least 1 encounter.

Table 2 details the nationwide coverage rates per disease module and encounter type as of September 18, 2023. It includes the cumulative target population size and the unique number of patients screened or monitored at least once. During this period, DMP screenings led to new diagnoses: 144,768 for hypertension, 490,346 for diabetes, 534,423 for high cardiovascular risk, and 3,545,573 for obesity. These individuals were diagnosed with these chronic diseases for the first time, following evidence-based clinical guidelines.

Age histograms of DMP patients who have been screened or monitored at least once are provided per sex in Figure 6.

Piloting studies occurred from October to December 2019, and the system has been fully operational nationwide since January 2020. Use notably increased with FMP salary incentivization calculations on July 1, 2021 (Figure 7), showing monthly encounter numbers by module from the start of 2021. Since then, encounters have steadily risen, with minor drops during summer holidays, and the distribution among DMP modules has remained consistent.

Figure 8 displays the distribution of total DMP encounters per city in Turkey, with colors intensifying as encounter numbers rise. Although higher numbers generally align with city populations, outliers exist, as seen in the top 10 performing cities outlined in Table 3. Despite Istanbul having Turkey's largest population, it only slightly surpasses Ankara in DMP encounters. This is mainly due to the high patient load per FMP in Istanbul. FMPs overseeing over 4000 citizens are exempt from DMP use due to their heavy workload. Table 3 also provides patient average age and encounter duration information.

The performance of the FMPs is assessed monthly. The cumulative targets and realized achievement rates for January 2023 are provided in Table 4. An achievement rate of 23.1% (4,508,841/19,546,041) for the entire population represents significant advancement compared with the 3.9% (511,198/13,117,900) achievement rate in July 2021.

The DMP system recommends personalized treatment goals such as systolic blood pressure, low-density lipoprotein cholesterol, and weight based on clinical guidelines. After a treatment goal is set, the DMP also assesses progress toward the goal in subsequent encounters. As of September 18, 2023, approximately 12.4 million of these treatment goals have been assessed, and the achievement rates are presented in Table 5. These assessments provide valuable information for FMPs caring for their patients.

At present, the performance of the FMPs is quantitatively calculated based on the number of performed encounters. However, the MOH envisions transitioning to a qualitative performance evaluation in midterm, where treatment goals and their achievement rates will play a significant role.

The system is highly performant and scalable. On a selected working day, February 14, 2023, the onFHIR.io HL7 FHIR Repository handled a total of 105.7 million FHIR interactions with an average response time of 31.3 milliseconds. During peak times of the day, the system can effortlessly manage up to 5000 FHIR interactions per second. Multimedia Appendix 2 illustrates the distribution and average response time of FHIR interactions on this day.

Among all FHIR requests, 57.4% (60.7 million) are search interactions, which are extensively used by the DMP web app to find, display, and forward specific clinical concept values to CDSS. Following search interactions, update interactions make up 33.2% (35.1 million) of the requests and are also used for resource creation when a provided resource ID is available. The average response times for read and search interactions are only 3.9 and 6.4 milliseconds, respectively. In the case of transactions and batch interactions, the average response times are even lower than update interaction alone, thanks to the parallelization of contained requests within onFHIR.io. As of September 18, 2023, onFHIR.io maintains a repository of 16.3 billion FHIR resources, totaling 22.4 terabytes in size, including care planning data by DMP and EHR/PHR data synchronized from e-Nabiz.

Table 1. Total screening and monitoring encounters per module.

Module	Screening (n=45,166,536), n (%)	Monitoring (n=28,548,733), n (%)	Total (n=73,715,269), n (%)
Hypertension	13,857,594 (30.7)	12,046,449 (42.2)	25,904,043 (35.1)
Obesity	18,029,994 (39.9)	800,480 (2.8)	18,830,474 (25.5)
Diabetes	8,914,193 (19.7)	5,071,646 (17.8)[a]	13,985,839 (19.0)
CVD[b] risk	4,364,755 (9.7)	9,182,814 (32.2)	13,547,569 (18.4)
Older adult	N/A[c]	1,447,344 (5.1)	1,447,344 (2.0)

[a]Only the patients monitored in primary care are listed; advanced obesity cases (a BMI over 40 kg/m^2 or a BMI between 30 and 40 kg/m^2 supported with additional comorbidities) are monitored in secondary and tertiary care.

[b]CVD: cardiovascular disease.

[c]N/A: not applicable.

Table 2. Coverage rate of citizens in target population lists.

Module and encounter type	All citizens in target population, n	Screened and monitored patients, n	Coverage rate (%)
Hypertension			
Screening	48,443,467	10,820,774	22.3
Monitoring	14,943,378	4,083,057	27.3
Obesity			
Screening	59,956,288	14,640,013	24.4
Monitoring	769,654[a]	383,920	49.9
Diabetes			
Screening	27,450,172	6,486,947	23.6
Monitoring	7,588,543	2,472,585	32.6
CVD[b] risk			
Screening	17,276,617	3,319,070	19.2
Monitoring	17,759,500	5,078,665	28.6
Older adult			
Monitoring	8,770,474	1,056,766	12.0

[a]Only those in the primary care obesity monitoring list, as explained in Table 1.

[b]CVD: cardiovascular disease.

Figure 6. Age histograms of disease management platform patients: female on the left and male on the right.

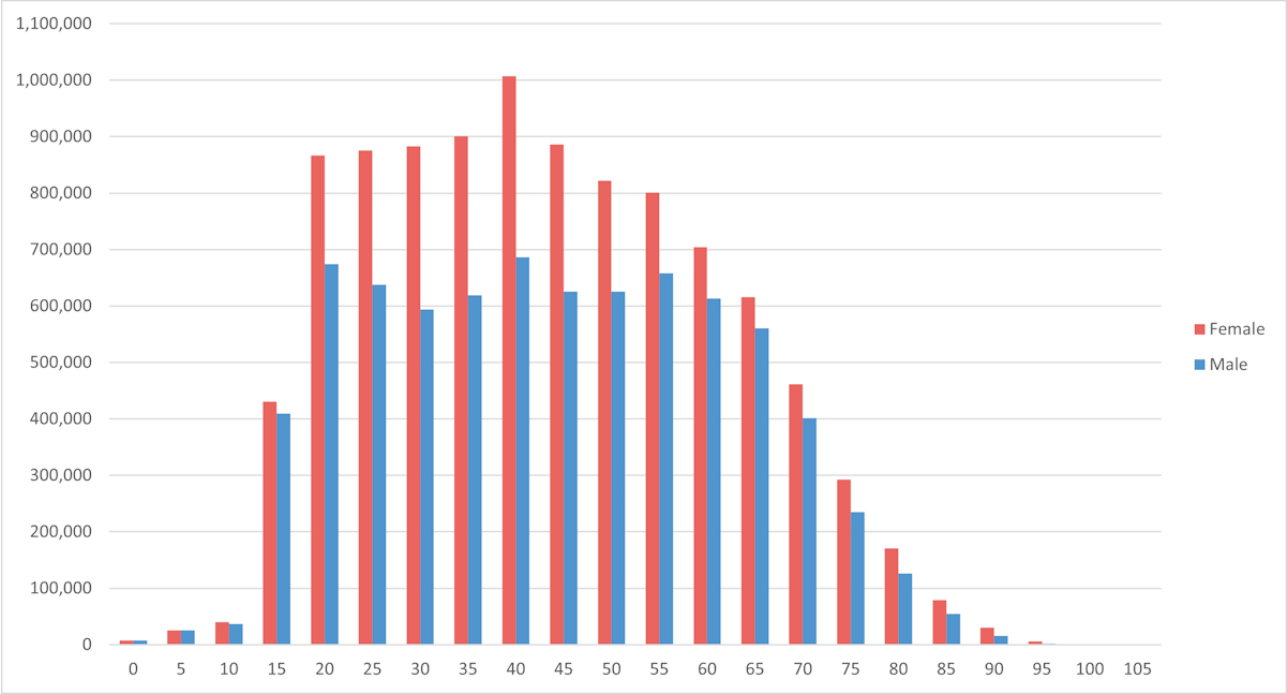


Figure 7. Disease management platform encounters per month by module. CVD: cardiovascular disease.

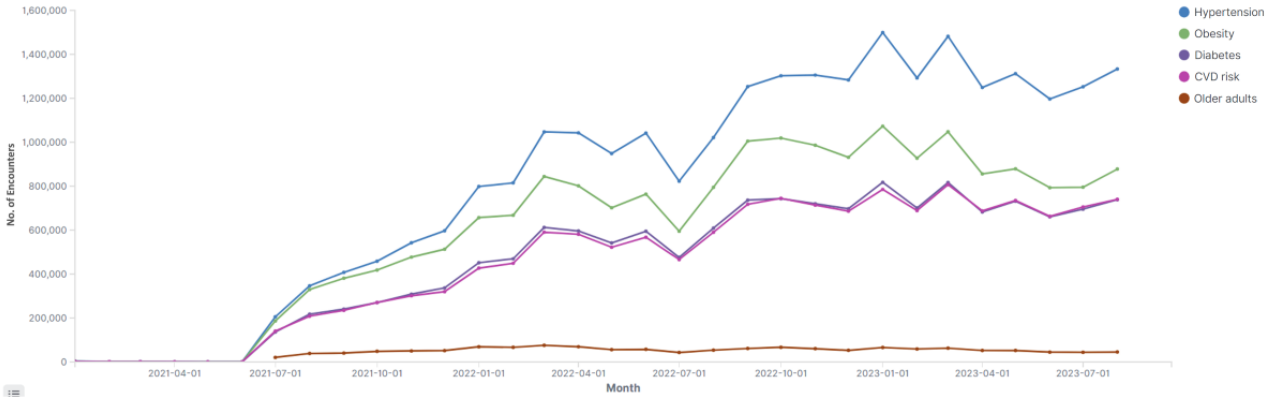


Figure 8. Encounters by city on a map.

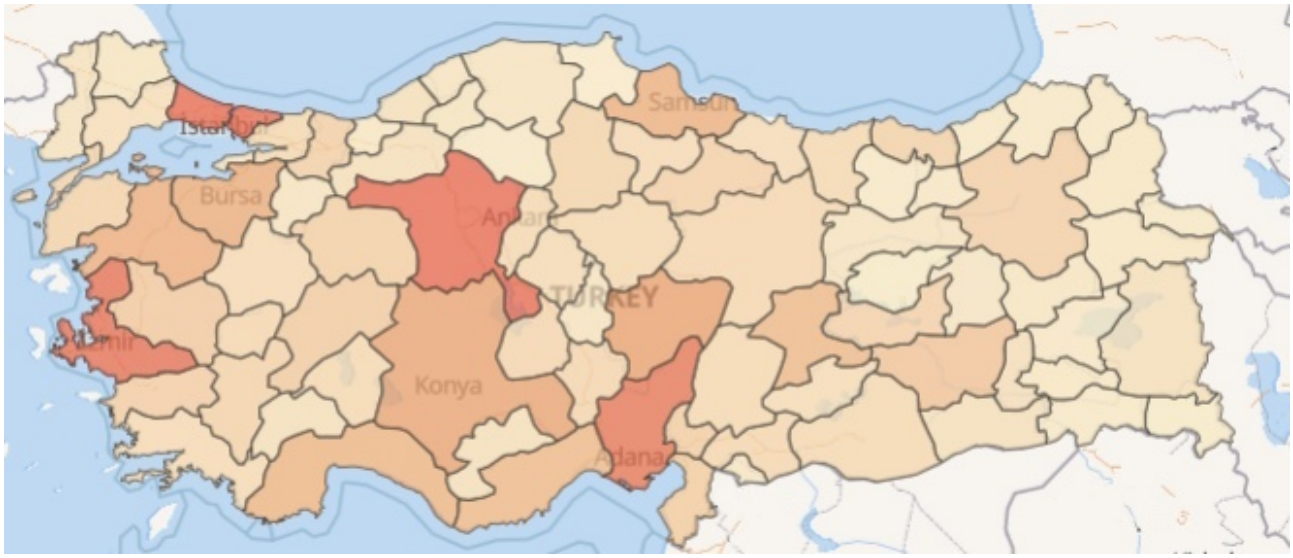


Table 3. Top 10 performing cities.

City	Total population[a]	Rank[b]	Number of encounters	Number of patients	Average age of patients (years)	Average duration (minutes)
Istanbul	15,907,951	1	4,974,972	1,286,640	50.5	1.21
Ankara	5,782,285	2	4,918,745	1,084,797	51.7	1.13
Izmir	4,462,056	3	4,395,653	937,607	53.4	1.18
Adana	2,274,106	7	3,531,441	709,075	51.3	0.99
Kayseri	1,441,523	15	2,633,349	479,849	51.7	1.06
Antalya	2,688,004	5	2,617,274	636,535	51.4	1.08
Konya	2,296,347	6	2,607,459	579,927	51.2	1.15
Bursa	3,194,720	4	2,408,817	525,016	52.3	1.18
Balikesir	1,257,590	17	2,322,111	425,171	55.7	1.13
Samsun	1,368,488	16	2,237,385	422,438	54.2	1.06

[a]2022 census data by the Turkish Statistical Institute (TurkStat).

[b]The rank of cities in Turkey by total population count.

Table 4. Screening and monitoring encounters per module in January 2023.

Module and encounter type	Monthly target	Number of encounters	Achievement rate (%)
Hypertension			
Screening	3,868,662	789,699	20.4
Monitoring	4,795,117	709,004	14.8
Obesity			
Screening	4,849,306	1,083,414	22.3
Monitoring	53,770	51,512	95.8
Diabetes			
Screening	898,665	670,502	74.6
Monitoring	2,128,955	279,071	13.1
CVD[a] risk			
Screening	742,634	262,419	35.3
Monitoring	1,488,004	589,523	39.6
Older adult			
Monitoring	720,928	73,697	10.2
Total	19,546,041	4,508,841	23.1

[a]CVD: cardiovascular disease.

Table 5. Achievement rates of treatment goals.

Treatment goal	Achievement rate (%)
Systolic BP[a]	88.8
Diastolic BP	94.2
Fasting glucose	52.0
HbA_{1c}[b]	61.5
LDL[c] cholesterol	14.8
HDL[d] cholesterol	63.2
Triglyceride	52.6
Weight	5.6
BMI	6.3
Waist circumference	2.9

[a]BP: blood pressure.

[b]HbA_{1c}: hemoglobin A_{1c}.

[c]LDL: low-density lipoprotein.

[d]HDL: high-density lipoprotein.

Discussion

Principle Findings and Lessons Learned

We have demonstrated that as of September 18, 2023, the DMP has been used by more than 25,000 users to conduct over 73 million screening and monitoring encounters for more than 16 million individuals. The national directive incentivizing FMPs to conduct screening and monitoring for chronic diseases is one of the contributing factors to this success.

We demonstrated the platform's efficient horizontal scalability, serving thousands of HCPs daily without performance issues. DMP screenings identified approximately 150,000 new hypertension cases, over 490,000 diabetes cases, more than 500,000 high cardiovascular risk cases, and over 3.5 million obesity cases. This allowed timely treatment in line with evidence-based guidelines.

We have shown that the system seamlessly interoperates with existing national EHR via HL7 FHIR. It enables accessing and processing patient data from various sources to provide personalized care plan guidance, maximizing the effectiveness

of evidence-based decision support services. The DMP has achieved all 5 levels of the 5S Model as proposed by Haynes [44] for the successful implementation of information services for evidence-based health care decisions. Continuous cocreation activity involving members of the Turkish MOH has contributed this success, along with the interoperability architecture based on international standards. On the other hand, we have collected feedback from FMPs to encourage us to also enable seamless integration with the national e-Prescription and national appointment system. FMPs need to manually input prescription and appointment recommendations into the other systems. Future plans include integrating these national systems directly to the DMP as well.

Although we have demonstrated that, through a repeatable and well-defined cocreation methodology, the system can be easily extended to address additional diseases, it still requires implementation effort from developers. We plan to extend the DMP system with administrative interfaces. This will enable subject matter experts from the MOH to create new disease screening and monitoring modules using form-based design interfaces.

Finally, although FMPs conduct screening and monitoring, specialists can view patient dashboards but cannot perform encounters; this can be easily enabled with the DMP's role-based access control mechanism, pending organizational decisions for national-scale implementation.

The system is operated as a part of national health IT ecosystem funded by the budget of the Turkish MOH. Open-source technologies have been used; hence, additional licensing fee has not incurred. Approximately 80% of the budget is spent for software development, 15% for project management, and 5% for training costs. Initial development phase has lasted 2 years. In the last 2.5 years, the system is under maintenance, and new disease modules have been developed.

Prospective Benefits and Impact

The system paves the way forward value-based care, where patient outcomes are monitored, and providers are incentivized for improving health. Currently, the DMP sets individual clinical goals (eg, HbA_{1c} and BMI) based on evidence-based guidelines. It monitors FMP performance in achieving these targets through close screening and monitoring. FMPs are presently incentivized based on screening and monitoring visits, but the system is ready to adopt value-based care by monitoring clinical targets.

DMP implementation opens opportunities to collect real-time research data, measuring the effectiveness of nationwide disease management protocols. Continuously gathering information about patients' disease status and recording outcomes from screening and monitoring visits, the generated data provide valuable insights into disease management.

Conclusions

This paper introduces a nationwide DMP designed for effective chronic disease screening and management, aligning with evidence-based clinical guidelines to enhance health care quality. With its user-friendly interfaces, it guides FMPs through personalized care planning with checklists for medication orders, referrals, lab tests, and risk screening. The system has been operational nationwide since January 2020. We have demonstrated seamless EHR integration, scalability, performance, and effectiveness in early diagnosis and meeting clinical targets. Future work includes a comprehensive study to analyze the direct clinical and cost-saving effects of the DMP on chronic disease management in Turkey.

Acknowledgments

The authors wish to acknowledge administrative and technical support by the following departments of the Turkish Ministry of Health: Department of Public Health Informatics, Department of Chronic Diseases and Elderly Health, Department of Healthy Nutrition and Active Life, Department of Data Management, and Department of Standards and Accreditation; Türksat, and Innova.

Data Availability

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

Authors' Contributions

GBLE and MMU carried out conceptualization of the paper; GBLE, MY, TN, SP, MG, YK, AAS, SG, AD, ZOA, and BE established the methodology of the study; GBLE, MY, TN, SP, MG, and YK developed the software; ZOA, BE, SA, MMU, and SB contributed validation studies; GBLE, MY, and TN carried out formal analysis; MY, TN, SP, MG, YK, AAS, and SG contributed to data curation; GBLE, MY, TN, SP, MG, and MMU wrote the manuscript; AAS, YK, SG, ZOA, BE, and SA reviewed and edited the manuscript; GBLE, MY, TN, SP, and MG created the visualizations in the manuscript; AD and SB supervised the study; GBLE, MY, BE, and MMU coordinated project administration. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Eligibility criteria for target populations for screening and monitoring encounters.
[DOCX File , 25 KB - medinform_v12i1e49986_app1.docx]

Multimedia Appendix 2
FHIR interactions in a day: request count on the left and average response time on the right. FHIR: Fast Healthcare Interoperability Resources.
[PNG File , 98 KB - medinform_v12i1e49986_app2.png]

References

1. Life tables, 2017-2019. Turkish Statistical Institute. 2020. URL: https://data.tuik.gov.tr/Bulten/Index?p=Hayat-Tablolari-2017-2019-33711 [accessed 2023-03-22]
2. Non-communicable diseases and risk factors cohort study for Turkey. Republic of Turkey Ministry of Health. URL: https://hsgm.saglik.gov.tr/depo/birimler/kronik-hastaliklar-ve-yasli-sagligi-db/Dokumanlar/Raporlar/v9s_NCDkohort_.pdf [accessed 2023-11-02]
3. Yucesan M, Gul M, Mete S, Celik E. A forecasting model for patient arrivals of an emergency department in healthcare management systems. In: Bouchemal N, editor. Intelligent Systems for Healthcare Management and Delivery. Hershey, Pennsylvania: IGI Global; 2019:266-284.
4. Omid P, Bilal A, Despotou G, Keung SNLC, Mohamad Y, Gappa H, et al. CAREPATH methodology for development of computer interpretable, integrated clinical guidelines. 2023 Presented at: DSAI 2022: 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion; August 31-September 2, 2022; Lisbon, Portugal. [doi: 10.1145/3563137.3563155]
5. Recommendations on digital interventions for health system strengthening–research considerations. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/WHO-RHR-19.9 [accessed 2023-11-02]
6. Agarwal S, Glenton C, Tamrat T, Henschke N, Maayan N, Fønhus MS, et al. Decision-support tools via mobile devices to improve quality of care in primary healthcare settings. Cochrane Database Syst Rev 2021;7(7):CD012944 [FREE Full text] [doi: 10.1002/14651858.CD012944.pub2] [Medline: 34314020]
7. 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;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
8. 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;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
9. Fortmann J, Lutz M, Spreckelsen C. System for context-specific visualization of clinical practice guidelines (GuLiNav): concept and software implementation. JMIR Form Res 2022;6(6):e28013 [FREE Full text] [doi: 10.2196/28013] [Medline: 35731571]
10. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293(10):1223-1238 [FREE Full text] [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
11. Shiffman RN, Liaw Y, Brandt CA, Corb GJ. Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. J Am Med Inform Assoc 1999;6(2):104-114 [FREE Full text] [doi: 10.1136/jamia.1999.0060104] [Medline: 10094063]
12. Poller L, Shiach CR, MacCallum PK, Johansen AM, Münster AM, Magalhães A, et al. Multicentre randomised study of computerised anticoagulant dosage. European concerted action on anticoagulation. Lancet 1998;352(9139):1505-1509 [FREE Full text] [doi: 10.1016/s0140-6736(98)04147-6] [Medline: 9820298]
13. Samore MH, Bateman K, Alder SC, Hannah E, Donnelly S, Stoddard GJ, et al. Clinical decision support and appropriateness of antimicrobial prescribing: a randomized trial. JAMA 2005;294(18):2305-2314 [FREE Full text] [doi: 10.1001/jama.294.18.2305] [Medline: 16278358]
14. Goud R, de Keizer NF, ter Riet G, Wyatt JC, Hasman A, Hellemans IM, et al. Effect of guideline based computerised decision support on decision making of multidisciplinary teams: cluster randomised trial in cardiac rehabilitation. BMJ 2009;338:b1440 [FREE Full text] [doi: 10.1136/bmj.b1440] [Medline: 19398471]
15. Filippi A, Sabatini A, Badioli L, Samani F, Mazzaglia G, Catapano A, et al. Effects of an automated electronic reminder in changing the antiplatelet drug-prescribing behavior among Italian general practitioners in diabetic patients: an intervention trial. Diabetes Care 2003;26(5):1497-1500 [FREE Full text] [doi: 10.2337/diacare.26.5.1497] [Medline: 12716811]
16. Erturkmen GBL, Yuksel M, Sarigul B, Arvanitis TN, Lindman P, Chen R, et al. A collaborative platform for management of chronic diseases via guideline-driven individualized care plans. Comput Struct Biotechnol J 2019;17:869-885 [FREE Full text] [doi: 10.1016/j.csbj.2019.06.003] [Medline: 31333814]

17. von Tottleben M, Grinyer K, Arfa A, Traore L, Verdoy D, Keung SNLC, et al. An integrated care platform system (C3-Cloud) for care planning, decision support, and empowerment of patients with multimorbidity: protocol for a technology trial. JMIR Res Protoc 2022;11(7):e21994 [FREE Full text] [doi: 10.2196/21994] [Medline: 35830239]
18. Lobach DF, Hammond WE. Computerized decision support based on a clinical practice guideline improves compliance with care standards. Am J Med 1997;102(1):89-98 [FREE Full text] [doi: 10.1016/s0002-9343(96)00382-8] [Medline: 9209205]
19. Dexter PR, Perkins S, Overhage JM, Maharry K, Kohler RB, McDonald CJ. A computerized reminder system to increase the use of preventive care for hospitalized patients. N Engl J Med 2001;345(13):965-970 [FREE Full text] [doi: 10.1056/NEJMsa010181] [Medline: 11575289]
20. Wang Z, An J, Lin H, Zhou J, Liu F, Chen J, et al. Pathway-driven coordinated telehealth system for management of patients with single or multiple chronic diseases in China: system development and retrospective study. JMIR Med Inform 2021;9(5):e27228 [FREE Full text] [doi: 10.2196/27228] [Medline: 33998999]
21. Ramirez M, Chen K, Follett RW, Mangione CM, Moreno G, Bell DS. Impact of a "chart closure" hard stop alert on prescribing for elevated blood pressures among patients with diabetes: quasi-experimental study. JMIR Med Inform 2020;8(4):e16421 [FREE Full text] [doi: 10.2196/16421] [Medline: 32301741]
22. Burack RC, Gimotty PA, Simon M, Moncrease A, Dews P. The effect of adding Pap smear information to a mammography reminder system in an HMO: results of randomized controlled trial. Prev Med 2003;36(5):547-554 [FREE Full text] [doi: 10.1016/s0091-7435(02)00062-2] [Medline: 12689799]
23. McPhee SJ, Bird JA, Fordham D, Rodnick JE, Osborn EH. Promoting cancer prevention activities by primary care physicians. Results of a randomized, controlled trial. JAMA 1991;266(4):538-544. [Medline: 2061981]
24. Bates DW, Kuperman GJ, Rittenberg E, Teich JM, Fiskio J, Ma'luf N, et al. A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests. Am J Med 1999;106(2):144-150. [doi: 10.1016/s0002-9343(98)00410-0] [Medline: 10230742]
25. Overhage JM, Tierney WM, Zhou XH, McDonald CJ. A randomized trial of "corollary orders" to prevent errors of omission. J Am Med Inform Assoc 1997;4(5):364-375 [FREE Full text] [doi: 10.1136/jamia.1997.0040364] [Medline: 9292842]
26. Wasylewicz ATM, Scheepers-Hoeks AMJW. Clinical decision support systems. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science. Cham (CH): Springer International Publishing; 2019.
27. Stagg BC, Stein JD, Medeiros FA, Wirostko B, Crandall A, Hartnett ME, et al. Special commentary: using clinical decision support systems to bring predictive models to the glaucoma clinic. Ophthalmol Glaucoma 2021;4(1):5-9 [FREE Full text] [doi: 10.1016/j.ogla.2020.08.006] [Medline: 32810611]
28. 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 [FREE Full text] [doi: 10.1136/jamia.2001.0080527] [Medline: 11687560]
29. 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 [accessed 2023-11-02]
30. Franck CP, 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;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
31. Welcome to FHIR. HL7 FHIR Release 4. URL: https://hl7.org/fhir/R4/ [accessed 2023-12-22]
32. HL7 FHIR Accelerator™ Program. HL7 International. URL: https://www.hl7.org/about/fhir-accelerator/ [accessed 2023-11-02]
33. What is FHIR? The Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/sites/default/files/2019-08/ONCFHIRFSWhatIsFHIR.pdf [accessed 2023-11-02]
34. Heat wave: the U.S. is poised to catch FHIR in 2019. HealthITBuzz. 2018. URL: https://www.healthit.gov/buzz-blog/interoperability/heat-wave-the-u-s-is-poised-to-catch-fhir-in-2019[accessed [accessed 2023-11-02]
35. 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]
36. 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]
37. FHIR (Fast Healthcare Interoperability Resources). NHS Digital. 2022. URL: https://digital.nhs.uk/services/fhir-apis [accessed 2023-11-02]
38. Medical Informatics Initiative demonstrates ability to standardise healthcare data across Germany according to FHIR. Medical Informatics Initiative Germany. 2019. URL: https://www.medizininformatik-initiative.de/en/medizininformatik-initiative-bundesweit-einheitliche-auswertbarkeit-von-versorgungsdaten [accessed 2023-11-02]
39. HL7 FHIR® based secure data repository. onFHIR.io. URL: https://onfhir.io/ [accessed 2023-11-02]
40. SMART App launch: scopes and launch context. HL7 International. URL: http://hl7.org/fhir/smart-app-launch/1.0.0/scopes-and-launch-context/index.html [accessed 2023-11-02]

41. Birinci S. National Healthcare Technology Initiative. In: Kacır MF, Seker M, Dogrul M, editors. National Technology Initiative. Ankara: Turkish Academy of Sciences Publication; 2022:305-328.
42. USS Services. URL: https://usskurumsal.saglik.gov.tr/kurumsalservisler/#HYP [accessed 2023-11-02]
43. HL7 CDS hooks. HL7 International. URL: https://cds-hooks.hl7.org/ [accessed 2023-11-02]
44. Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. Evid Based Med 2006;11(6):162-164 [FREE Full text] [doi: 10.1136/ebm.11.6.162-a] [Medline: 17213159]

Abbreviations

CDSS: clinical decision support service
DMP: Disease Management Platform
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
FMP: family medicine practitioner
HbA_{1c}: hemoglobin A_{1c}
HCP: health care professional
HL7: Health Level Seven
ICD-10: International Classification of Diseases, Tenth Revision
MOH: Ministry of Health
PHR: personal health record

Edited by C Perrin; submitted 16.06.23; peer-reviewed by J Galvez-Olortegui, J Pevnick; comments to author 13.09.23; revised version received 21.09.23; accepted 29.11.23; published 19.01.24.

Please cite as:
Ulgu MM, Laleci Erturkmen GB, Yuksel M, Namli T, Postacı Ş, Gencturk M, Kabak Y, Sinaci AA, Gonul S, Dogac A, Özkan Altunay Z, Ekinci B, Aydin S, Birinci S
A Nationwide Chronic Disease Management Solution via Clinical Decision Support Services: Software Development and Real-Life Implementation Report
JMIR Med Inform 2024;12:e49986
URL: https://medinform.jmir.org/2024/1/e49986
doi:10.2196/49986
PMID:38241077

Implementation Report

Ten Years of Experience With a Telemedicine Platform Dedicated to Health Care Personnel: Implementation Report

Claudio Azzolini[1,2,3], MD; Elias Premi[3,4], MD; Simone Donati[3,5], MD; Andrea Falco[2,6], MEng; Aldo Torreggiani[7], MEng; Francesco Sicurello[3,8], BSc; Andreina Baj[5], MD; Lorenzo Azzi[5], MD; Alessandro Orro[2,3,8], MEng; Giovanni Porta[5], MD; Giovanna Azzolini[2], BSc; Marco Sorrentino[2,9], JD; Paolo Melillo[10], MEng; Francesco Testa[10], MD; Francesca Simonelli[10], MD; Gianfranco Giardina[11], MEng; Umberto Paolucci[12], MEng

[1]Advisory Council of e-Health and Telemedicine, University of Insubria of Varese-Como, Varese, Italy
[2]TM95 Srl, Milan, Italy
[3]Italian Association of Telemedicine and Medical Informatics, Milan, Italy
[4]Department of Life Sciences and Biotechnologies, University of Insubria, Varese-Como, Italy
[5]Department of Medicine and Surgery, University of Insubria, Varese-Como, Italy
[6]Alfa Design Studio, Milan, Italy
[7]T&C Srl, Milan, Italy
[8]Institute of Biomedical Technologies, National Research Council, Milan, Italy
[9]Bms Farm Law Firm, Milan, Italy
[10]Multidisciplinary Department of Medical, Surgical and Dental Sciences, University of Campania Luigi Vanvitelli, Naples, Italy
[11]DDay.it, Milan, Italy
[12]Up Invest Srl, Milan, Italy

Corresponding Author:
Claudio Azzolini, MD
Advisory Council of e-Health and Telemedicine
University of Insubria of Varese-Como
Via Guicciardini 9
Varese, 21100
Italy
Phone: 39 0332393603
Email: claudio.azzolini@uninsubria.it

Abstract

Background: Telemedicine, a term that encompasses several applications and tasks, generally involves the remote management and treatment of patients by physicians. It is known as transversal telemedicine when practiced among health care professionals (HCPs).

Objective: We describe the experience of implementing our telemedicine Eumeda platform for HCPs over the last 10 years.

Methods: A web-based informatics platform was developed that had continuously updated hypertext created using advanced technology and the following features: security, data insertion, dedicated software for image analysis, and the ability to export data for statistical surveys. Customizable files called “modules” were designed and built for different fields of medicine, mainly in the ophthalmology subspecialty. Each module was used by HCPs with different authorization profiles.

Implementation (Results): Twelve representative modules for different projects are presented in this manuscript. These modules evolved over time, with varying degrees of interconnectivity, including the participation of a number of centers in 19 cities across Italy. The number of HCP operators involved in each single module ranged from 6 to 114 (average 21.8, SD 28.5). Data related to 2574 participants were inserted across all the modules. The average percentage of completed text/image fields in the 12 modules was 65.7%. All modules were evaluated in terms of access, acceptability, and medical efficacy. In their final evaluation, the participants judged the modules to be useful and efficient for clinical use.

Conclusions: Our results demonstrate the usefulness of the telemedicine platform for HCPs in terms of improved knowledge in medicine, patient care, scientific research, teaching, and the choice of therapies. It would be useful to start similar projects across various health care fields, considering that in the near future medicine as we know it will completely change.

(JMIR Med Inform 2024;12:e42847) doi:10.2196/42847

KEYWORDS

telemedicine; ophthalmology; eHealth; informatics platform; health care professional; patient care; information technology; data warehouse

Introduction

Context

Medicine has typically involved physicians engaging face to face with patients. However, many teleconsultation projects have now been developed, particularly during the COVID-19 pandemic era, which has boosted teleconsultations in all medical specialties [1-4].

Alongside telemedicine between physicians and patients, there is also transversal telemedicine, which is conducted between health care professionals (HCPs). Our experience with this topic started in 1996 and has demonstrated the feasibility of training young ophthalmic vitreoretinal surgeons working in nonoptimal environments (postwar Bosnia), using telemedicine (via a satellite link) in Milan and Sarajevo [5,6]. Input from the above experiences [7,8] constituted the basis for our understanding of the needs of HCPs and the developmental direction of the dedicated telemedicine platform, giving users access, with appropriate personal authorization, from anywhere and at any time.

Problem Statement

The problem to be solved is the difficulty of sharing patients' clinical data and images among health personnel for efficient evaluation. This process should be multidisciplinary, involving actors such as physicians from different specialties, nurses, technicians, orthoptists, geneticists, residents, and tutors who need access to a common database holding key patient information.

Similar Interventions

Our scientific literature analysis identified a number of publications about implementation projects involving telemedicine platforms. These projects were mainly based on COVID-19 management and aimed to support different systems to provide health care in emergency conditions [9-11]. The purpose of these initiatives is to foster telecare and telemonitoring and to reduce the need for patients to visit hospitals or medical centers [12-17]. Our program is oriented in a different direction: the Eumeda web-based medical platform was developed for sharing patients' medical data among physicians. The platform has expanded its services to many HCPs. This paper describes how database modules for the clinical databank and trials, as well as second opinion services, were created and have now been implemented.

Methods

Aims and Objectives

The aim of this implementation program was to broaden the applications of our telemedicine platform with a transversal approach targeted at health care personnel. This process took place over the last 10 years with the creation of different projects aimed at clinical data collection, teleconsultation, and gathering second opinions. Various modules have been built for the platform (Textbox 1) for use by HCPs at different times. Twelve representative modules for different clinical projects are described in Table 1 [18-23].

We identified outcome measures and evaluated overall parameters for access, acceptability, and medical efficacy of the platform (Textbox 2).

Textbox 1. Building a module in 8 steps. The time required for the final release varies between modules (from 1 to 3 months for more complex ones). The original source code for the modules created belongs to the medical platform.

1. Initial agreement between the entity applying the module (university, company, institution, or representative association) and the manager of the medical platform (MP)

2. Signing of detailed operational form (with project requirements, such as the type of project, number of health care professionals and structures involved, and the importance of images) by the main users of the module and the scientific coordinator (who has knowledge of medicine planning and the potentiality and limits of medical informatics) of the MP

3. "Shoulder-to-shoulder" work by the scientific coordinator of the MP and main team programmer of the MP

4. Development of alpha software (not yet stable and still incomplete) to be shown to the entity that will use the module for changes and additions

5. Development of beta software with almost all functionalities

6. Massive data entry by the MP programmer to find bugs or software incompatibilities

7. Completion of beta software with automatic control functionalities (eg, alert icons to prevent inappropriate data from being entered, numerical limitations, and priorities to be respected in data entry or blocking of inappropriate saving) for users to check and identify any small changes required

8. Release of final version in a meeting with users, with explanatory text embedded in the module

Table 1. The left-hand column lists the 12 representative projects for which many modules have been built for the medical platform. The modules designed have been managed by health care personnel over the last 10 years in different locations in Italy.

Module	Description	Module type	Purpose	Timeframe of project activity	Holder	Sponsor
1	Teleconsultation in retinal diseases [18]	Second opinions[a]	Feasibility of second opinions among physicians	1 Month (during 2011)	Insubria University, Varese-Como	Comed Research nonprofit association, Milan
2	Age-related maculopathy [19]	Group[b] (10 locations)	Acceleration of anti–vascular endothelial growth factor therapy	19 Months (2011-2012)	T&C Srl, Milan, Italy	Novartis Pharma SpA, Origgio, Italy
3	Retinal pathology samples and correlated genes [20]	Data[c]	Collection of data on gene expression	4 Months (2012-2013)	Insubria University, Varese-Como	Insubria University, Varese-Como
4	Epiretinal macular membrane [21]	Data	Collection of data on disease morphology and functionality	10 Months (2015-2016)	Insubria University, Varese-Como	Insubria University, Varese-Como
5	Inherited eye diseases	Data	Collection of data on genetic eye diseases	2017-present	Ophthalmological Unit II, University of Naples	Ophthalmological Unit II, University of Naples; Rome Foundation
6	Retinal dystrophy due to rare *RPE65* gene mutation [22]	Group (9 locations)	Collection of data on disease	16 Months (2018-2020)	Ophthalmological Unit II, University of Naples	Retina Italia nonprofit association, Milan
7	Second opinions among resident physicians	Second opinions	Feasibility of second opinions in didactics	4 Months (during 2019)	Comed Research nonprofit association, Milan	Bayer Italy SpA, Milan
8	Instrumental data in multiple sclerosis	Group (2 locations)	Collection of multidisciplinary data on disease	2019-present	Neurological Unit, Insubria, University Varese-Como	Insubria University, Varese-Como
9	Epidemiological data on COVID-19 in workers	Group (2 locations)	Search for COVID-19 in throat, saliva, and tears	3 Months (during 2020)	SEA Company, Milan Linate-Malpensa Airports	SEA Company, Milan Linate-Milan Malpensa Airports
10	SARS-CoV-2 on throat and ocular surfaces [23]	Group (2 locations)	Search for SARS-CoV-2 in throat and tears in COVID-19 patients	1 Month (during 2020)	T&C Srl, Milan. Italy	Insubria University, Varese-Como
11	Potential malignant oral lesions	Group (2 locations)	Collection of data on disease	2021-present	Orthodontics Unit, Insubria University, Varese-Como	Insubria University, Varese-Como
12	Maculopathies and anti-aging medicine	Data	Collection of data on diseases and follow-up	2022-present	Claude Boscher, MD	Claude Boscher, MD

[a]Second opinions: second opinions from health care professionals at the same or a different institution.

[b]Group: shared database used by health care professionals at more than one institution.

[c]Data: shared database used by health care professionals at a single institution.

Textbox 2. Result options for the questionnaire for each health care professional, with relative scores. The final score is given by the sum of the partial scores (maximum 9, minimum 3). Scores equal to or higher than 6 are considered to indicate approval.

Access to the network by computer or mobile devices

- Poor: score of 1
- Good: score of 2
- Very good: score of 3

Acceptability of the procedures

- Poor: score of 1
- Good: score of 2
- Very good: score of 3

Medical efficacy

- Poor: score of 1
- Good: score of 2
- Very good: score of 3

Blueprint Summary

Design of Key Features and Roadmap

The design of the implementation program was oriented to develop three types of operational modules, integrated with one another where necessary: (1) a databank of diseases for clinical or scientific studies, (2) a database for groups of HCPs in different locations, giving them access to shared data from trial studies, and (3) a functionality enabling physicians to seek second opinions. The key points of the implemented modules were easy accessibility, complete acceptability for HCPs, data reliability, and overall medical efficacy considering all health specialties. The roadmap followed these principles and several new projects involving HCPs produced specific modules, which were created for the platform and take advantage of its benefits as a whole.

Technological Design and Infrastructure

Since 2010, the Eumeda platform has used continuously updated versions of PHP, an HTML-embedded web scripting language built to a high standard using advanced technology [24], which has the advantage of speed, flexibility, low use of resources, and compatibility with all web servers. PHP does not require a high level of machine resources to run and is therefore very fast and lends itself to applications with external integration.

Main Features of the Platform

Information technology services can be accessed via monitors or mobile devices and include current advanced technologies, such as the following: 24-7, 365-day-a-year access, easy data image insertion in electronic medical records, image comparison and overlapping, and SMS and email notification, when necessary, for fast interactivity.

Customizable Modules

The platform includes customizable files called “modules” that are designed and built for each project according to its needs in collaboration with professionals from different knowledge areas (Textbox 1). Each module functions to support the features and advantages of the entire platform. No data are sent directly to or from HCPs’ hardware. HCPs are able to see data in the central database, accessing this information remotely. All HCPs have a personal access code depending on their authorization level, enabling them to view, insert, or modify data in specific fields, close the electronic medical record (EMR) data temporarily or permanently, and export data for statistical surveys. The platform allows for individual and group interaction among HCPs at different sites. A remote “prompt assistance” service is provided for each module when necessary.

Module Functionality

Several functions can be activated, with open pop-ups showing the rationale of each study, its population, the provenance of resources, and the operating HCPs with various authorization profiles. The data entry procedure is quick, intuitive (Figure 1), and guided by many system alerts in the case of errors. When necessary, warning notifications are sent to users via SMS or email. Special software can be created, if requested, to support HCPs’ data evaluation and clinical decisions [25-27] (Figure 2). Data extraction for statistical surveys is immediate (Figure 1). At the end of each study period, the HCPs evaluated the project using a 3-point scoring system (Textbox 2).

Figure 1. Example of the main tasks and procedures for a module (module 6 in Table 1) on the medical platform: (1) entry to the system by the health care professional with their personal access key, after which they select the modules that they are qualified and authorized to use; (2) access to a list of operative centers with their own lists of patients and respective electronic medical records relating to the first and follow-up visits; (3) individual patient electronic medical record folder, which allows for the easy and quick insertion of data and multiple images at any time, as well as access to successive masks (a repository of images is available that allows image overlapping and comparison; Figure 3); (4) quick data extraction for statistical purposes.

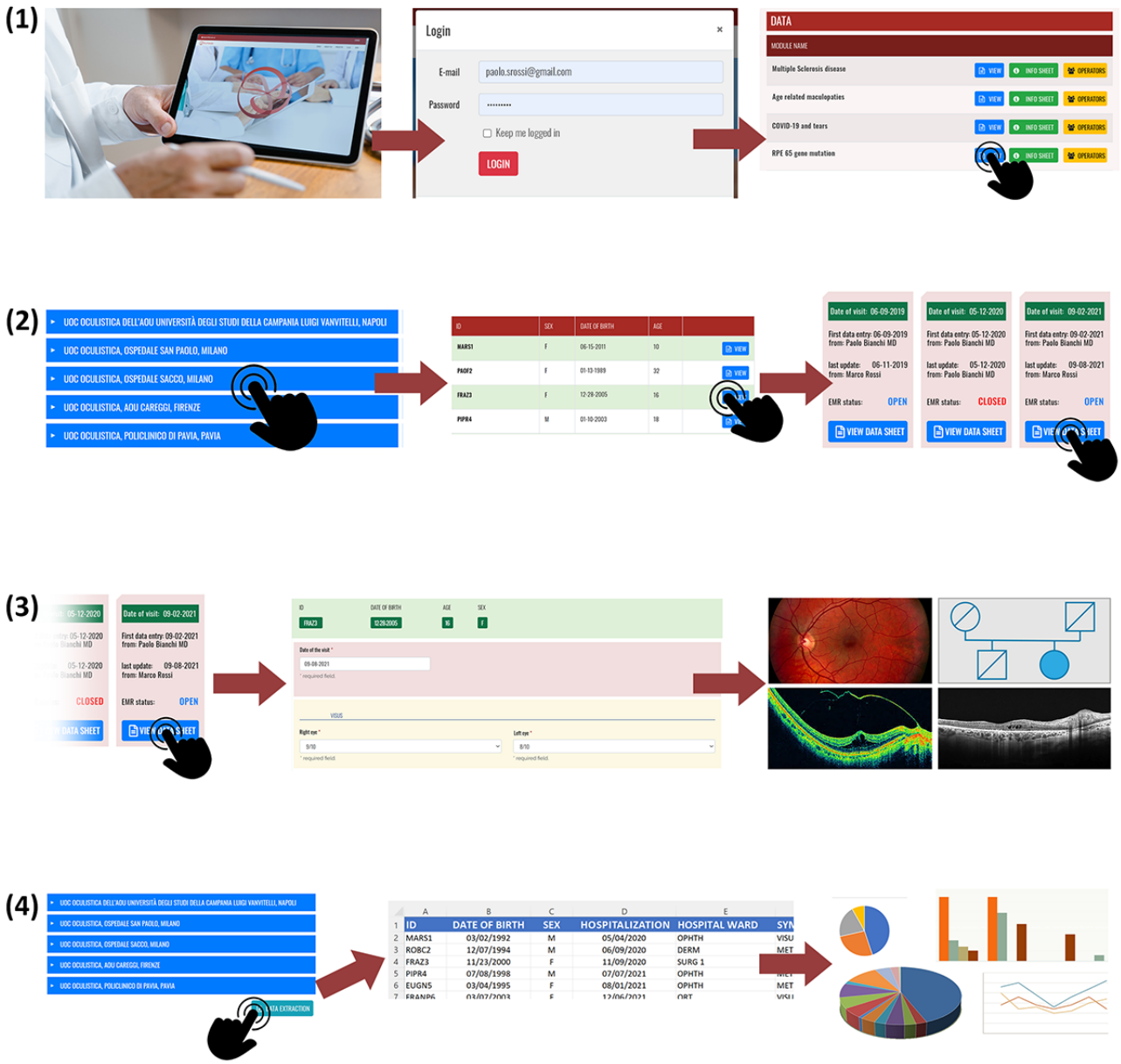
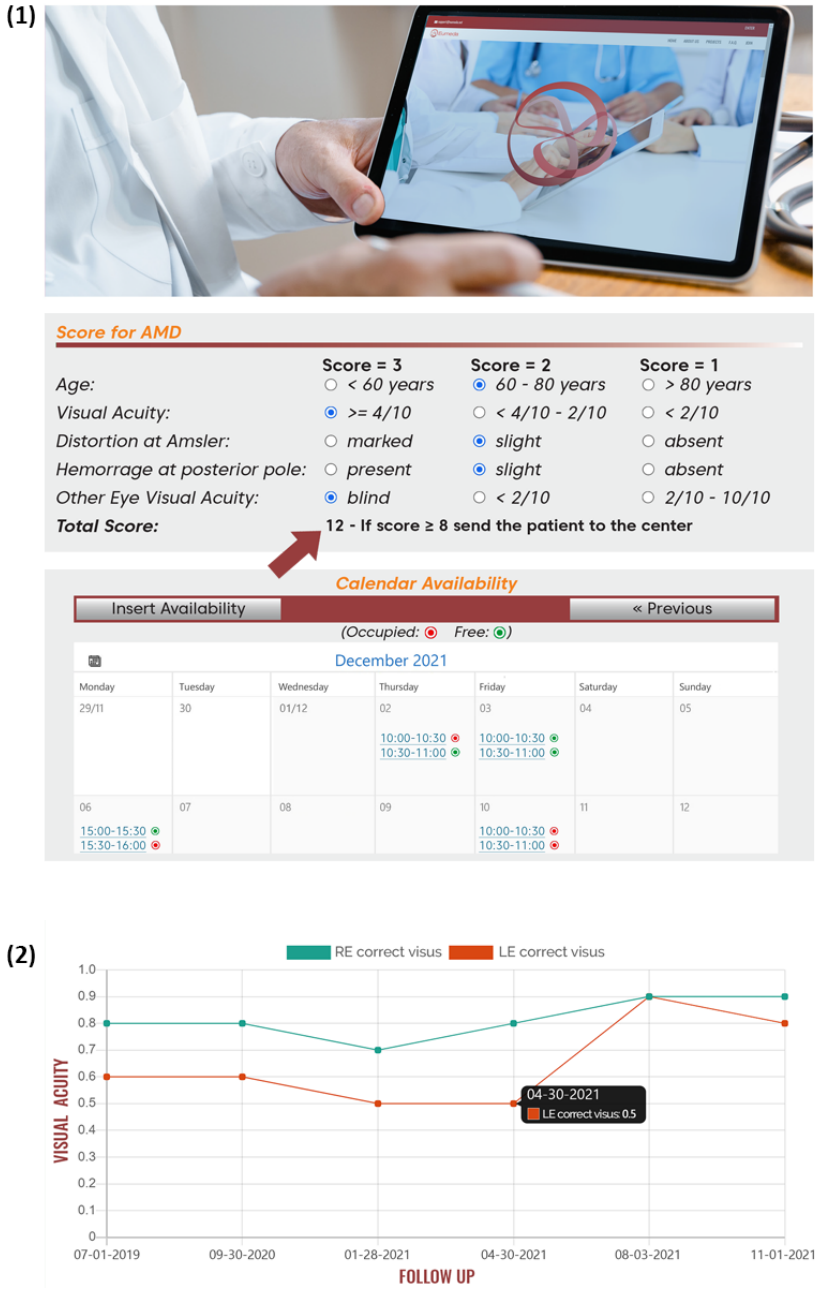


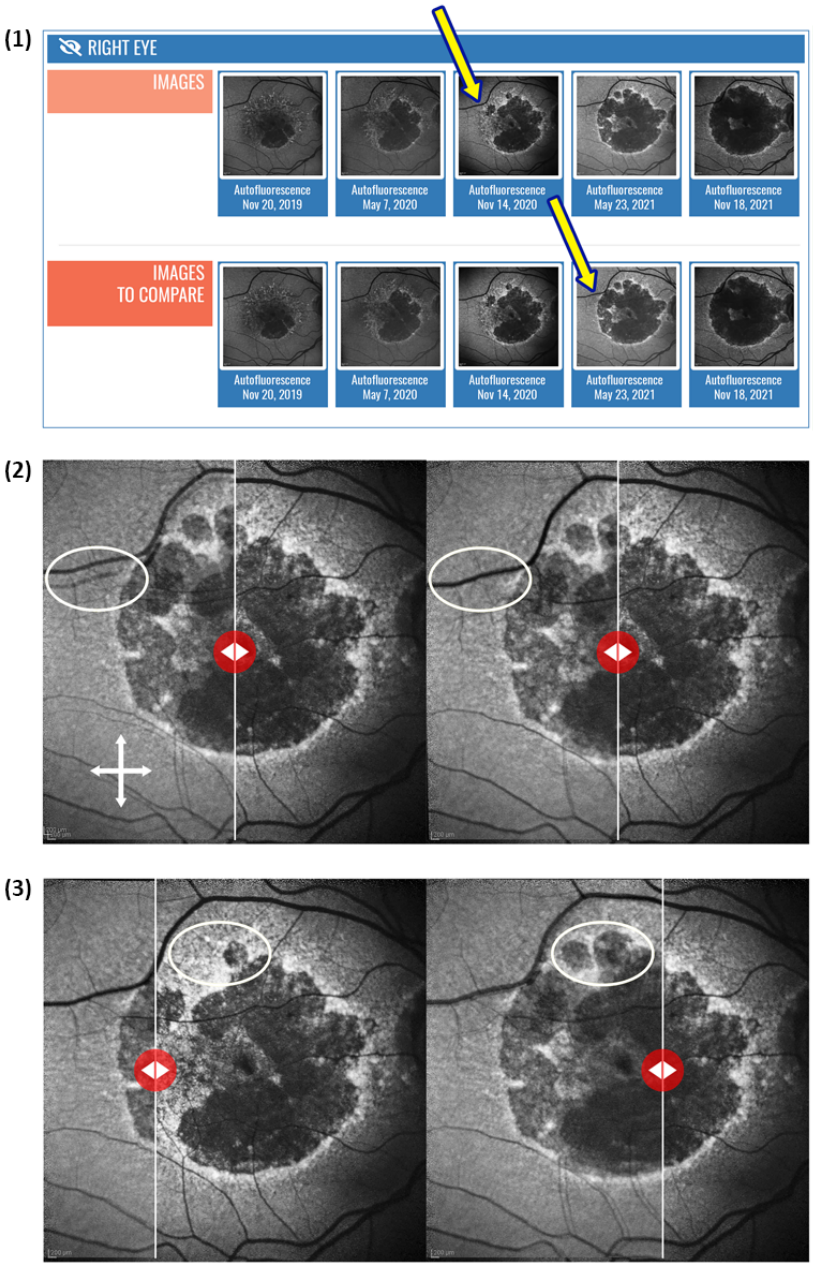
Figure 2. Examples of 2 special software programs designed to support health care professional activity. (1) For medical care decisions, each diagnostic variable of a disease is given a numeric value, and the software automatically provides a total score (shown by the arrow). If the value exceeds a defined score, the software advises general physicians to send the patient to an appropriate center at the next available appointment (module 2 in Table 1). (2) For tracking patients' clinical course, visual acuity data (or any other numerical data) are inserted into a patient's electronic medical record and the graph is updated in real time. The health care professional can see at a glance the functional course of the disease. RE: right eye; LE: left eye.



Images

Dedicated software allows the uploading of even high-resolution images and videos in a few seconds. A shared whiteboard for all images is available for each module. Image magnification and comparison software enables morphological changes to be observed over time in detail (Figure 3).

Figure 3. Example of image comparison: (1) selection of 2 images from a patient's electronic medical record (in this case, the patient had degenerative retinal maculopathy) uploaded at the 6-month follow-up (shown by the arrows); (2) creation of an overlap image that can be adjusted by clicking and moving the white cross; a special transparency application allows the images to be accurately superimposed on each other (shown by the white rings); (3) evaluation of morphological changes in the disease over time (within the white rings) by moving the overlap line back and forth (shown by the white line) using the red button.



Type of Technology

The Eumeda software platform is closed-source and owned by a private company that grants access through contracts. The platform was developed with the Wappler (Wappler.io) integrated development environment.

Targets

The target user base includes physicians of different specialties, nurses, technicians, orthoptists, geneticists, residents, and tutors, all of whom rely on access to a common database holding key patient information.

The target sites are hospitals, private offices, and medical hub centers working with spoke-peripheral centers. We involved medical structures equipped with technology and staff prepared to use hardware and software (Figure 4).

Figure 4. Locations of health care professionals (round dots) in Italy who have used the platform over the years (the map shows the Italian names of the cities). The Supervisory Center (SC) in Milan has responsibility for the data warehouse and help desk as well as a general coordination role.

Data

Data Location

All data were uploaded and stored in a data warehouse in Milan, Italy (Datasys Srl from 2001 to 2012; then Aruba Business Srl, provided by IRQ10 Srl, from 2013 until the present), to ensure data security and uninterrupted availability. Automated daily backups are a security measure guarding against the loss of data.

Data Entry Policy

HCPs must agree to the liability agreement, ownership agreement, and a code of conduct before using the platform. In all modules, data entry is performed in accordance with the guidelines of the Declaration of Helsinki and its subsequent revisions [28]. Informed consent forms are collected by the health facilities. In cases where data analysis included a therapeutic choice, approval from the relevant ethics committee or a qualified local committee was obtained, as in modules 6 and 11 in Table 1.

Data Security and Privacy

Data security and privacy are guaranteed by the latest generation of servers with secure backups. Data are protected on several levels: (1) individual HCPs receive access keys generated by the system; (2) subjects' personal data are encrypted and stored in a separate table in the cloud; (3) the system binds clinical data to personal data only when accessing hardware with a special algorithm; and (4) if necessary, a ready-to-use informed consent form can be downloaded for signature.

Responsibility for and Ownership of Data

According to European Union (EU) and Italian rules, liability for entered medical data, including cloud storage, lies with the HCP entering such data (acting as the "controller," as clearly defined in EU Regulation 679/16) and the manager of the medical platform and server farm (acting as "processors," as clearly defined in EU Regulation 679/16). Ownership of the data, including the purpose and methods for processing the data, belongs to the entity applying the module, who acts as the "controller."

Interoperability

The Eumeda software platform is not accessible from specific application programming interface clients by users, so it does not use data standards such as Health Level Seven. However, it implements the International Classification of Diseases, 10th

revision, system internally to classify pathologies. Clinical imaging is managed with current standard protocols.

Participating Entities

A nonprofit organization (Comed Research) initially implemented (from 2001) the projects. Subsequently, a joint venture between 2 for-profit companies (T&C Srl and TM95 Srl) managed the platform. These partners supply hardware, create software, or participate as web designers, hosting companies, or law firms. The funders of the implementation projects are public universities, hospitals and foundations, nonprofit medical associations, private companies, and physicians working in private offices (Table 1).

The society that created the main platform is the owner of all the implementations. The entities that applied the modules hold the ownership and the intellectual property.

Budget Planning

The total budget covered different phases according to implementation progression and project type for a period of 1 to 3 years. The costs included preliminary planning and the final draft (up to 30%), programming for final front view on the computer screen (30%), and user training (10%), as well as the pilot phase (5%), operation (10%), initial service (10%), and ongoing reports (5%). Selected projects could be conducted for free based on their importance or visibility for the platform.

Sustainability

The projects were initially funded (from 2001) by a nonprofit organization (Comed Research), which relies on donations from companies or nonprofit medical associations. Since January 1, 2017, the platform has been managed by a joint venture between 2 for-profit companies. The business model is based on the type and duration of the projects developed during the implementation phase, financed by different entities. The end of the project foresees the dissemination of the results with potential permanent effects.

Implementation (Results)

Coverage

The projects developed during the implementation phase have national coverage, encompassing a large number of Italian regions and their referent hospitals. The developed modules evolved over time, with varying degrees of interconnectivity, in different centers in 19 cities across Italy (Figure 4). In 2 modules (modules 1 and 2 in Table 1), HCPs from the referring regional areas were closely involved. The number of HCPs (at different levels) using individual modules ranged from 6 to 114 (average 21.8, SD 28.5).

Outcomes

Implementing the telemedicine platform allowed us to build several modules that could be used by HCPs at different times. The characteristics of 12 representative modules used over the last 10 years for different clinical projects are shown in Table 1.

Over time, our experience has led us to concentrate on three types of operational modules, integrated with one another if necessary: (1) a databank of diseases for clinical or scientific studies (eg, module 4; Table 1), (2) a database for groups of HCPs in different locations, giving them access to shared data (eg, module 6; Table 1), and (3) a functionality enabling physicians to seek second opinions (eg, module 7; Table 1).

The overall outcomes are reported in Table 2. Up to now, more than 250 HCPs have used the platform for several effective and operational projects. The total number of participants inserted in the modules is 2574. The percentage of data entered in the text or image fields for each module ranged from 20% to 95% (with an average of 65.7%). The evaluation score for each module was calculated as the sum of 3 partial scores (Textbox 2): out of all the modules, the first (module 1, the first to be created) was the one with the lowest evaluation score (Table 2). The average number of requests for technological support varied from 5 per month (in the case of simpler modules) to 9 (for more complex ones).

Table 2. Results pertaining to the designed modules shown in Table 1.

Module	Description	Centers involved, n	HCPs[a] involved, n	Participants whose data were inserted, n	Text/image fields for each EMR[b], n (fields that were filled in, %)	Beneficial effects	Evaluation score[c] (minimum positive score)
1	Teleconsultation in retinal diseases [18]	1 Retina center, 17 territorial offices	18	52	30 (60)	Useful teleconsultation among doctors	109 (108)
2	Age-related maculopathy [19]	11 Retina centers	114	678	65 (85)	Improvements in patients' functional final outcomes	803 (684)
3	Retinal pathology samples and correlated genes [20]	1 Ophthalmological center, 1 genetic center	11	12	65 (80)	Better understanding of molecular mechanisms	Not acquired
4	Epiretinal macular membrane [21]	2 Ophthalmologic centers, 1 human anatomy center	11	28	25 (65)	Identification of ultramicroscopic features of membranes	80 (66)
5	Inherited eye diseases[d]	1 Ophthalmological center, 1 genetic center	14	1145[e]	480[e] (20)	Increased knowledge of genetic eye diseases	In progress
6	Retinal dystrophy due to rare *RPE65* gene mutation [22]	9 Retinal-genetic centers	28	60	260 (65)	Identification of suitable patients for therapy	200 (168)
7	Second opinions among resident physicians[d]	4 University ophthalmological departments	19	110	12 (85)	Resident physicians' learning accelerated	140 (114)
8	Instrumental data in multiple sclerosis[d]	2 Neurological centers, 2 ophthalmological centers	6	58[e]	450[e] (18)	Recognition of the disease in the subclinical stage	In progress
9	Epidemiological data on COVID-19 in workers[d]	2 Care offices at 2 airports	9	298	30 (90)	Collection of useful diagnostic data on COVID-19 and how the disease is transmitted	75 (54)
10	SARS-CoV-2 on throat and ocular surfaces [23]	14 Medical units	20	108	34 (95)	Increased knowledge of COVID-19	165 (120)
11	Potential malignant oral lesions	4 Medical units	6	15[e]	50[e] (68)	Better prevention and therapy	In progress
12	Maculopathies and anti-aging medicine	1 Retina center	6	10[e]	110[e] (58)	Significantly better care	In progress

[a]HCP: health care professional (physicians from different specialties, nurses, technicians, orthoptists, geneticists, residents, tutors [employees were excluded]).

[b]EMR: electronic medical record (for each patient, considering first visit and all follow-ups).

[c]Sum of 3 partial scores for access, acceptability, and medical efficacy at end of the active working period (described in Textbox 2).

[d]Unpublished data.

[e]At the time of writing this paper.

Clinical fallout can be identified more easily with the use of this telemedicine platform because of the visibility of a database shared by HCPs (modules 3, 4, 5, 8, and 9; Table 2). Furthermore, data on rare diseases (collected from a large number of centers) can be used to identify patients who would benefit from expensive new therapies (module 6; Table 2). By sharing medical data, physicians and residents can learn better and faster (modules 1 and 7; Table 2), and the possibility of having a databank helps them to discover potential, as yet unknown disease complications (module 10; Table 2). Patient follow-up with dedicated software helps HCPs to locate better treatment options, identify preventive interventions (modules 2 , 11 and 12; Table 2) and track patients and their outcomes in real time.

Lessons Learned

Our program has multiple success factors that may be considered in future implementations or in the creation of similar telemedicine platforms and modules. First, the technological infrastructure of the platform is modern, highly versatile, and continuously updated by technical staff. The use of the latest generation of servers with secure, daily backups guarantees that no data loss occurs, while data security and privacy are protected on several levels, as specified in the Methods section. Second, data entry and retrieval in each module are immediate. Each module has different blocks of information that are well separated, including an explanation of the rationale of the study and practical guidance on how to insert data, as well as different HCP access profiles, patient IDs, EMRs, images, and statistical surveys. Third, no images are transmitted among HCPs. All images are stored on the main server and are viewed remotely without any deterioration. A dedicated procedure even allows the insertion of high-resolution images (through common connection links) immediately or very quickly. Rapid viewing is greatly appreciated by users, in addition to the possibility of enlarging, comparing, and superimposing, as well as being able to see in detail the morphological changes, even minimal ones, of a pathology over time (Figure 3). Lastly, different authorization profiles are given to HCPs, which enables them to access modules on the central server once they have agreed to abide by the terms of the liability and ownership agreements and the code of conduct, using personal passwords to view, change, or modify data and images. Module coordinators usually have total control of their respective modules and can compile statistical surveys using all the data, while other HCPs may only be able to enter data and images in accordance with their remit and authorized access level. A great amount of work has been undertaken to ensure that the user-friendly platform is up and running. A remote service is available by mail or telephone. All modules are visible both from monitors and mobile devices. In particular, the second opinion module may be suitable for use with mobile health (mHealth).

We consider the following points more as challenges than limits to implementation. The construction phase of each module is of critical importance, and a single medical interlocutor must be the voice of all HCPs (Textbox 1). The main mandatory factors involved in building a module include a scientific coordinator as the central figure and the participation of someone with both medical and IT skills. Finally, older HCPs tended to struggle with working on the platform, while the younger operators adapted quickly, were not disconcerted by the technology, and showed interest and satisfaction with the projects they carried out [29-34].

The presumed budgets of each project, divided into direct (eg, coordinators, IT programmers, law firms) and indirect (eg, travel, equipment, insurance) costs have been considered in the final balance.

The following recommendations may assist in overcoming many barriers to telemedicine practice among HCPs. First, the amount of preparatory work needed (Textbox 1) tends to be underestimated. Second, it is difficult to create systems for sharing text and images with appropriate levels of usability. Third, bureaucracy is often an obstacle, and self-regulation codes in telemedicine need official authorization. Fourth, a suitable “network culture” is still lacking in medicine, due to multiple technical and human factors. The success of telemedicine among HCPs requires participation, responsibility, and a desire for effective collaboration to develop knowledge for the benefit of professionals and patients.

Discussion

Principal Findings

The technological infrastructure of telemedicine intended strictly for HCPs is specific to this field, is not easy to implement, and must be customized for each individual project. Key persons such as scientific coordinators (with specific knowledge of medicine and IT) and program managers must be well chosen for projects to succeed. The results of our projects have shown a range of benefits, including increased medical efficacy and clinical knowledge, improved patient care, enhanced teaching and integration among hospitals, and a more effective choice of therapies. It is necessary for work to be carried out on organizational, bureaucratic, and network culture issues where these are not yet fully accepted and on sustainable business plans.

We identified some difficulties and limitations in our implementation project that may also be considered useful for future or similar telemedicine projects. Building a module in the absence of straightforward ideas forced us to make major changes during construction, meaning that the preliminary work completed had to be discarded and redone. All the software involved in the platform modules must be customized according to the needs of the HCPs, which requires time and hard work. Too many text or image fields to fill out and include in EMRs make the system difficult to use and produce a very large final database that is not fully used (as happened with module 5).

Conclusion

In conclusion, our experience was that both physicians and patients were always satisfied to be part of this “community of health” supported by groups of HCPs working for their benefit and making them feel cared for. The detailed description of our implementation program may be useful to shorten the learning curve for others seeking to implement similar projects in many fields of medicine, which must be able to adapt to the continuously changing nature of medicine now and in the future.

Acknowledgments

We would like to thank the hundreds of health care professionals who have used and contributed over time to the development of the platform. Special thanks go to all engineers, programmers, and other personnel who in their various capacities have worked to improve the platform over the last 10 years, especially Francesco Oggionni (IT professional) and Valerio Tartaglia (IT professional). We are grateful to Ferruccio Fazio, MD, former Italian minister of health, and Gianfranco Ferla, MD, for their support and advice. We would also like to thank Roberta Romagnolo (Lexikon) for editing the manuscript.

Data Availability

The data on which this manuscript is based are available upon request to the corresponding author.

Conflicts of Interest

None declared.

References

1. Ebbert JO, Ramar P, Tulledge-Scheitel SM, Njeru JW, Rosedahl JK, Roellinger D, et al. Patient preferences for telehealth services in a large multispecialty practice. J Telemed Telecare 2023 May;29(4):298-303. [doi: 10.1177/1357633X20980302] [Medline: 33461397]
2. Schulz T, Long K, Kanhutu K, Bayrak I, Johnson D, Fazio T. Telehealth during the coronavirus disease 2019 pandemic: Rapid expansion of telehealth outpatient use during a pandemic is possible if the programme is previously established. J Telemed Telecare 2022 Jul;28(6):445-451 [FREE Full text] [doi: 10.1177/1357633X20942045] [Medline: 32686556]
3. Reitzle L, Schmidt C, Färber F, Huebl L, Wieler LH, Ziese T, et al. Perceived access to health care services and relevance of telemedicine during the COVID-19 pandemic in Germany. Int J Environ Res Public Health 2021 Jul 19;18(14):7661 [FREE Full text] [doi: 10.3390/ijerph18147661] [Medline: 34300110]
4. Capusan KY, Fenster T. Patient satisfaction with telehealth during the COVID-19 pandemic in a pediatric pulmonary clinic. J Pediatr Health Care 2021;35(6):587-591 [FREE Full text] [doi: 10.1016/j.pedhc.2021.07.014] [Medline: 34417077]
5. Azzolini C, Fontanella G, Mason A. A pilot study to train vitreoretinal surgeons by telemedicine. 1997 Presented at: Association for Research in Vision and Ophthalmology (ARVO) Annual Meeting; May 11-16; Fort Lauderdale, FL p. 397.
6. Karcić S, Azzolini C, Alikadić-Husović A. [Telemedicine in vitreoretinal surgery]. Med Arh 1999;53(3 Suppl 3):73-75. [Medline: 10870633]
7. Contini F, Prati M, Donati S, Azzolini C. Idiopathic macular hole: Multicentric clinical trial. 2006 Presented at: Association for Research in Vision and Ophthalmology Meeting; May; Fort Lauderdale, FL URL: https://iovs.arvojournals.org/article.aspx?articleid=2391305&resultClick=1
8. Mason A, Feliciani F, Morelli P. The Italian telemedicine SHARED project. 1998 Presented at: American Telemedicine Association Third Annual Conference; Orlando, FL p. 5.
9. Li S, Wang C, Lu W, Lin Y, Yen DC. Design and implementation of a telecare information platform. J Med Syst 2012 Jun;36(3):1629-1650. [doi: 10.1007/s10916-010-9625-6] [Medline: 21120592]
10. Clin L, Leitritz MA, Dietter J, Dynowski M, Burgert O, Ueffing M, et al. Design, implementation and operation of a reading center platform for clinical studies. Stud Health Technol Inform 2017;235:33-37. [doi: 10.3233/978-1-61499-753-5-33] [Medline: 28423750]
11. Lopez E, Berlin M, Stein R, Cozzi E, Bermudez A, Mandirola Brieux H, et al. Results of the use of the teleconsultation platform after 2 months of implementation. Stud Health Technol Inform 2020 Jun 16;270:1377-1378. [doi: 10.3233/SHTI200450] [Medline: 32570667]
12. Hasson SP, Waissengrin B, Shachar E, Hodruj M, Fayngor R, Brezis M, et al. Rapid implementation of telemedicine during the COVID-19 pandemic: Perspectives and preferences of patients with cancer. Oncologist 2021 Apr;26(4):e679-e685 [FREE Full text] [doi: 10.1002/onco.13676] [Medline: 33453121]
13. Beltrán V, von Marttens A, Acuña-Mardones P, Sanzana-Luengo C, Rueda-Velásquez SJ, Alvarado E, et al. Implementation of a teledentistry platform for dental emergencies for the elderly in the context of the COVID-19 pandemic in Chile. Biomed Res Int ;2022:6889285 [FREE Full text] [doi: 10.1155/2022/6889285] [Medline: 35330690]
14. Espay AJ, Hausdorff JM, Sánchez-Ferro Á, Klucken J, Merola A, Bonato P, et al. A roadmap for implementation of patient-centered digital outcome measures in Parkinson's disease obtained using mobile health technologies. Mov Disord 2019 May;34(5):657-663 [FREE Full text] [doi: 10.1002/mds.27671] [Medline: 30901495]
15. Nadar M, Jouvet P, Tucci M, Toledano B, Cyr M, Sicotte C. The implementation of a synchronous telemedicine platform linking off-site pediatric intensivists and on-site fellows in a pediatric intensive care unit: A feasibility study. Int J Med Inform 2019 Sep;129:219-225. [doi: 10.1016/j.ijmedinf.2019.06.009] [Medline: 31445259]

16. Kern C, Fu DJ, Kortuem K, Huemer J, Barker D, Davis A, et al. Implementation of a cloud-based referral platform in ophthalmology: making telemedicine services a reality in eye care. Br J Ophthalmol 2020 Mar;104(3):312-317 [FREE Full text] [doi: 10.1136/bjophthalmol-2019-314161] [Medline: 31320383]
17. Brenner B, Brancolini S, Eshraghi Y, Guirguis M, Durbhakula S, Provenzano D, et al. Telemedicine implementation in pain medicine: A survey evaluation of pain medicine practices in spring 2020. Pain Physician 2022 Aug;25(5):387-390 [FREE Full text] [Medline: 35901479]
18. Azzolini C. A pilot teleconsultation network for retinal diseases in ophthalmology. J Telemed Telecare 2011;17(1):20-24. [doi: 10.1258/jtt.2010.100305] [Medline: 21097561]
19. Azzolini C, Torreggiani A, Eandi C, Donati S, Oum MA, Vinciguerra R, et al. A teleconsultation network improves the efficacy of anti-VEGF therapy in retinal diseases. J Telemed Telecare 2013 Dec;19(8):437-442. [doi: 10.1177/1357633X13501760] [Medline: 24162839]
20. Azzolini C, Pagani IS, Pirrone C, Borroni D, Donati S, Al Oum M, et al. Expression of VEGF-A, Otx homeobox and p53 family genes in proliferative vitreoretinopathy. Mediators Inflamm 2013:857380 [FREE Full text] [doi: 10.1155/2013/857380] [Medline: 24227910]
21. Azzolini C, Congiu T, Donati S, Passi A, Basso P, Piantanida E, et al. Multilayer microstructure of idiopathic epiretinal macular membranes. Eur J Ophthalmol 2017 Nov 08;27(6):762-768. [doi: 10.5301/ejo.5000982] [Medline: 28525683]
22. Testa F, Murro V, Signorini S, Colombo L, Iarossi G, Parmeggiani F, et al. RPE65-associated retinopathies in the Italian population: a longitudinal natural history study. Invest Ophthalmol Vis Sci 2022 Feb 01;63(2):13 [FREE Full text] [doi: 10.1167/iovs.63.2.13] [Medline: 35129589]
23. Azzolini C, Donati S, Premi E, Baj A, Siracusa C, Genoni A, et al. SARS-CoV-2 on ocular surfaces in a cohort of patients with COVID-19 from the Lombardy Region, Italy. JAMA Ophthalmol 2021 Sep 01;139(9):956-963 [FREE Full text] [doi: 10.1001/jamaophthalmol.2020.5464] [Medline: 33662099]
24. PHP Manual. The PHP Group. 1997. URL: https://www.php.net/download-docs.php [accessed 2023-12-20]
25. Campbell JP, Lee AY, Abràmoff M, Keane PA, Ting DS, Lum F, et al. Reporting guidelines for artificial intelligence in medical research. Ophthalmology 2020 Dec;127(12):1596-1599 [FREE Full text] [doi: 10.1016/j.ophtha.2020.09.009] [Medline: 32920029]
26. Azzolini C, Donati S, Falco A. The digital citizen: duties and rights to build a fairer future society. 2022 Presented at: XXII Infopoverty World Conference; December 1; New York, NY.
27. Azzolini C, Donati S. The digital era: new horizons in medicine and rehabilitation. 2023 Presented at: XXIII National Congress of the Italian Association of Telemedicine and Medical Informatics; November 24-25; Rome, Italy.
28. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013 Nov 27;310(20):2191-2194. [doi: 10.1001/jama.2013.281053] [Medline: 24141714]
29. Pit SW, Velovski S, Cockrell K, Bailey J. A qualitative exploration of medical students' placement experiences with telehealth during COVID-19 and recommendations to prepare our future medical workforce. BMC Med Educ 2021 Aug 16;21(1):431 [FREE Full text] [doi: 10.1186/s12909-021-02719-3] [Medline: 34399758]
30. Mahabamunuge J, Farmer L, Pessolano J, Lakhi N. Implementation and assessment of a novel telehealth education curriculum for undergraduate medical students. J Adv Med Educ Prof 2021 Jul;9(3):127-135 [FREE Full text] [doi: 10.30476/jamp.2021.89447.1375] [Medline: 34277843]
31. Curioso WH, Peña-Ayudante WR, Oscuvilca-Tapia E. COVID-19 reveals the urgent need to strengthen nursing informatics competencies: a view from Peru. Inform Health Soc Care 2021 Sep 02;46(3):229-233. [doi: 10.1080/17538157.2021.1941974] [Medline: 34292802]
32. Bell KA, Porter C, Woods AD, Akkurt ZM, Feldman SR. Impact of the COVID-19 pandemic on dermatology departments' support of medical students: A survey study. Dermatol Online J 2021 Jul 15;27(7):14 [FREE Full text] [doi: 10.5070/D327754376] [Medline: 34391338]
33. Coe TM, McBroom TJ, Brownlee SA, Regan K, Bartels S, Saillant N, et al. Medical students and patients benefit from virtual non-medical interactions due to COVID-19. J Med Educ Curric Dev 2021;8:23821205211028343 [FREE Full text] [doi: 10.1177/23821205211028343] [Medline: 34368454]
34. Frankl S, Joshi A, Onorato S, Jawahir GL, Pelletier SR, Dalrymple JL, et al. Preparing future doctors for telemedicine: an asynchronous curriculum for medical students implemented during the COVID-19 pandemic. Acad Med 2021 Dec 01;96(12):1696-1701 [FREE Full text] [doi: 10.1097/ACM.0000000000004260] [Medline: 34323861]

Abbreviations

EU: European Union
EMR: electronic medical record
HCP: health care professional
mHealth: mobile health

Edited by C Perrin; submitted 13.10.22; peer-reviewed by J Gurp, van, M Venturini, S Pesälä, V Ramos; comments to author 27.01.23; revised version received 13.07.23; accepted 29.11.23; published 26.01.24.

Please cite as:
Azzolini C, Premi E, Donati S, Falco A, Torreggiani A, Sicurello F, Baj A, Azzi L, Orro A, Porta G, Azzolini G, Sorrentino M, Melillo P, Testa F, Simonelli F, Giardina G, Paolucci U
Ten Years of Experience With a Telemedicine Platform Dedicated to Health Care Personnel: Implementation Report
JMIR Med Inform 2024;12:e42847
URL: https://medinform.jmir.org/2024/1/e42847
doi:10.2196/42847
PMID:38277199

Implementation Report

Learnings From Implementation of Technology-Enabled Mental Health Interventions in India: Implementation Report

Sudha Kallakuri[1], MSc; Sridevi Gara[1], BE; Mahesh Godi[1], BE; Sandhya Kanaka Yatirajula[1], PhD; Srilatha Paslawar[1], MPH; Mercian Daniel[1], PhD; David Peiris[2,3], MBBS, MIPH, PhD; Pallab Kumar Maulik[1,3,4,5,6], MSc, MD, PhD

[1]George Institute for Global Health, New Delhi, India

[2]George Institute for Global Health, Sydney, Australia

[3]Faculty of Medicine, University of New South Wales, Sydney, Australia

[4]Department of Brain Sciences, Imperial College London, London, United Kingdom

[5]Prasanna School of Public Health, Manipal Academy of Higher Education, Manipal, India

[6]George Institute for Global Health, London, United Kingdom

Corresponding Author:
Sudha Kallakuri, MSc
George Institute for Global Health
308 Elegance Tower, Third Floor
Plot No 8, Jasola District Centre
New Delhi, 110025
India
Phone: 91 11 4158 8091
Email: skallakuri1@georgeinstitute.org.in

Abstract

Background: Recent years have witnessed an increase in the use of technology-enabled interventions for delivering mental health care in different settings. Technological solutions have been advocated to increase access to care, especially in primary health care settings in low- and middle-income countries, to facilitate task-sharing given the lack of trained mental health professionals.

Objective: This report describes the experiences and challenges faced during the development and implementation of technology-enabled interventions for mental health among adults and adolescents in rural and urban settings of India.

Methods: A detailed overview of the technological frameworks used in various studies, including the Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health pilot study, SMART Mental Health cluster randomized controlled trial, and Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums (ARTEMIS) study, is provided. This includes the mobile apps that were used to collect data and the use of the database to store the data that were collected. Based on the experiences faced, the technological enhancements and adaptations made at the mobile app and database levels are described in detail.

Implementation (Results): Development of descriptive analytics at the database level; enabling offline and online data storage modalities; customizing the Open Medical Record System platform to suit the study requirements; modifying the encryption settings, thereby making the system more secure; and merging different apps for simultaneous data collection were some of the enhancements made across different projects.

Conclusions: Technology-enabled interventions prove to be a useful solution to cater to large populations in low-resource settings. The development of mobile apps is subject to the context and the area where they would be implemented. This paper outlines the need for careful testing using an iterative process that may support future research using similar technology.

Trial Registration: SMART Mental Health trial: Clinical Trial Registry India CTRI/2018/08/015355; https://ctri.nic.in/Clinicaltrials/pmaindet2.php?EncHid=MjMyNTQ=&Enc=&userName=CTRI/2018/08/015355. ARTEMIS trial: Clinical Trial Registry India CTRI/2022/02/040307; https://ctri.nic.in/Clinicaltrials/pmaindet2.php?EncHid=NDcxMTE=&Enc=&userName=CTRI/2022/02/040307

(JMIR Med Inform 2024;12:e47504) doi:10.2196/47504

KEYWORDS

mental health; technological interventions; digital health; community intervention; implementation; eHealth; India; Asia; development; health technology

Introduction

The burden of mental disorders [1] and the treatment gap due to untreated mental disorders in low- and middle-income countries (LMICs) such as India is estimated to range between 75% and 85% [2], with 1 in every 27 individuals being treated for depression [3]. Technological solutions have been advocated to increase access to care, especially in primary health care settings in LMICs, to facilitate task-sharing, given the lack of trained mental health professionals. Research has indicated the effectiveness of employing technologies for addressing complex health concerns among people with mental illnesses. However, the cost-effectiveness of technology-enabled interventions compared to in-person interventions has not yet been established [4].

Technology-enabled service delivery models have increased access to care and facilitated service monitoring, with mobile health (mHealth) being one such strategy. The World Health Organization (WHO) defines mHealth as "a medical and public health practice that is supported by mobile devices, such as mobile phones, patient monitoring devices, and other wireless devices" [5]. mHealth in the form of electronic decision support systems (EDSSs) has been widely adopted by service users and providers for monitoring health status and for diagnosing and managing a range of health conditions, including mental disorders and substance use [6]. mHealth use has increased with increasing penetration of mobile network connectivity [7].

This paper highlights the processes involved in the development and implementation of technology-enabled interventions employed in three projects across rural and urban settings in India among adults and adolescents: the Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health (SMH) Pilot project [8], and two cluster randomized controlled trials (cRCTs), SMH trial [9] and the Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums (ARTEMIS) trial [10].

All three projects used EDSSs to facilitate the identification, diagnosis, and management of common mental disorders (CMDs), including depression, anxiety, psychological distress, and increased suicide risk. A task-sharing approach was used where nonphysician health workers known as Accredited Social Health Activists (ASHAs) and primary care doctors worked together to support people at high risk of CMDs [8-10].

This implementation report describes the experiences of using technology in implementing these three mental health projects, following implementation reporting guidelines [11].

Methods

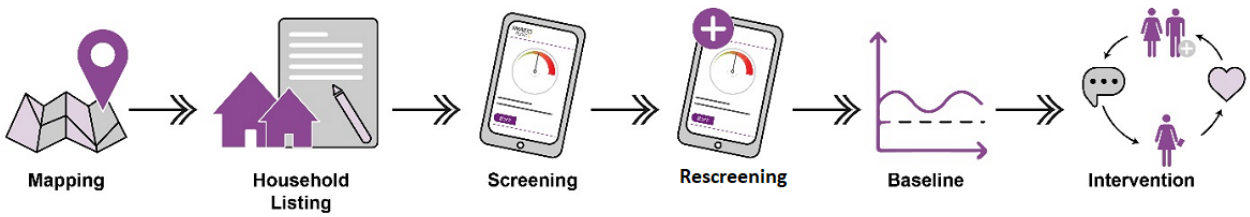
Aims and Objectives

This paper highlights the processes involved in the development and implementation of technology-enabled interventions employed in three projects across rural and urban settings among adults and adolescents in India.

Blueprint Summary

The overall technological framework of the SMH pilot study has two main components: a mobile app and a database. Different mobile apps were developed to collect data at divergent phases of the study (Figure 1). All apps were installed on 7-inch Android tablets for use by ASHAs/community women volunteers (CWVs), or primary health center (PHC) doctors. ASHAs are local women trained from the community with 8th-10th–grade education levels to support the implementation of health programs. While ASHAs work contractually, they are incentivized for their involvement in other projects. CWVs are women who reside in the same community where the study is being done. These CWVs were chosen from the slums and would have similar education level as ASHAs. They were trained on basic knowledge about mental health, along with the stigma and care of individuals with stress, depression, and increased suicide risk.

Figure 1. Different phases of the study.



The three studies underwent a formative phase, testing study tools and mobile apps while gauging user acceptance [12,13]. Iterations were made based on user feedback before the intervention phase. The technical team assessed the app and released a test version for research team testing. Once confirmed, a definitive version was used for data collection.

In the preintervention phase, geographical mapping and demarcation of the village boundaries were performed, followed by house listing to obtain accurate census data. Custom apps were developed for each step, including population screening for identifying individuals at risk of CMDs, which involved data collectors and ASHAs using specific screening tools. After screening, baseline data on various variables were collected before the intervention was implemented (Figure 1).

Technical Framework Design

The key components of the EDSS included the ASHA app, doctors app, and priority listing app (Table 1). Each ASHA had a finite set of individuals who lived in the geographical location covered by her. The tablets had encrypted, password-protected individual logins unique for every ASHA. Individuals screening positive were referred to primary care doctors for clinical diagnosis and treatment based on predetermined cut-off scores. The doctors used the WHO mhGAP-IG tool (version 1.0) [14] for diagnosing and treating people with CMDs, offering algorithm-based diagnoses and evidence-based treatment recommendations, including comorbidities. Doctors followed these recommendations, entering the type of care provided (pharmacological, psychological, referral, or combinations thereof) into their app. Doctors input the data to generate a traffic light–coded priorities list for ASHAs, indicating the status of screen-positive individuals in their area. Using color coding due to the low education levels of ASHAs, the list included pertinent questions on treatment adherence, social support, and stressors for each color category. The list was dynamic, changing based on doctors' updates during patient follow-up visits.

Table 1. Details of the apps used for the three studies and the target of the intervention.

App	Phase of the study	Users
SMART MH[a] (Pilot) and SMART MH trial focused on rural adults		
Listing app	Listing (household census data collection)	Data collectors
Screening app	Household screening for common mental disorders	ASHAs[b]
Baseline data collection app	Baseline: collected data on different variables and stressors triggering anxiety/depression	Data collectors
Intervention (ASHA app)	Intervention: for regular follow up of adults at high risk of CMDs[c] who sought care from the doctor or have yet to seek care	ASHAs
Intervention (doctor app)	Intervention: diagnosis and treatment for CMDs among adults	Primary care doctors
3M, 6M, and 12M app	Assessments at 3, 6, and 12 months of the intervention	Data collectors
ARTEMIS[d] trial focused on adolescents		
Listing app	Listing (household census data collection)	Data collectors
Screening app	Household screening for common mental disorders	Data collectors
Baseline data collection app	Baseline: collected data on different variables and stressors triggering anxiety/depression	Data collectors
Intervention (ASHA app)	Intervention: for regular follow up of adolescents who are at high risk of CMDs who sought care from the doctor or have yet to seek care	ASHAs
Intervention (doctor app)	Intervention: diagnosis and treatment for CMDs	Primary care doctors
3M, 6M, and 12 M app	Assessments at 3, 6, and 12 months of the intervention	Data collectors

[a]SMART MH: Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health.

[b]ASHA: Accredited Social Health Activist.

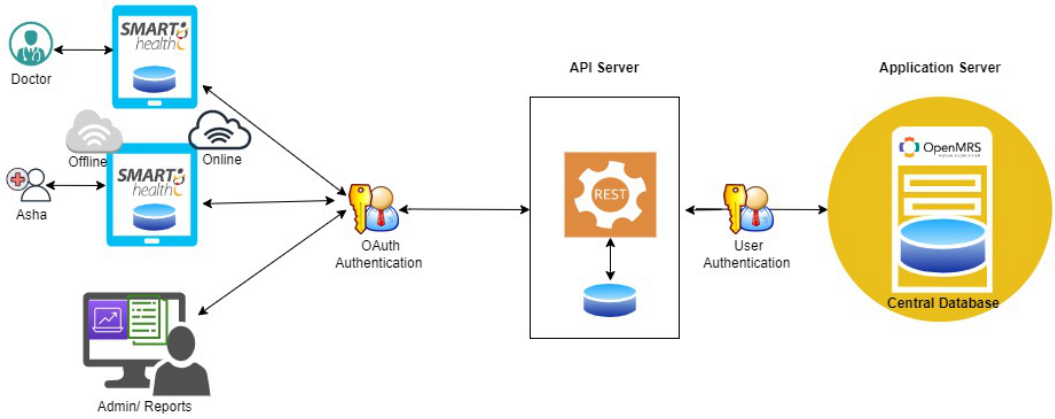
[c]CMD: common mental disorder.

[d]ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums.

Identifying an Electronic Medical Record System

All three projects utilized apps based on the Open Medical Record System (OpenMRS) [15], a community-driven open-source software for medical record storage and processing. OpenMRS is robust, scalable for large interventions, and customizable to study workflows and data collection needs. OpenMRS was chosen for these projects as it is freely available. Based on our earlier experience, the functionalities were suitable for our mental health projects [16]. Data collected on tablets underwent authentication and were transferred to the application programming interface (API) server, which were then sent to the application server housing the central OpenMRS database (Figure 2).

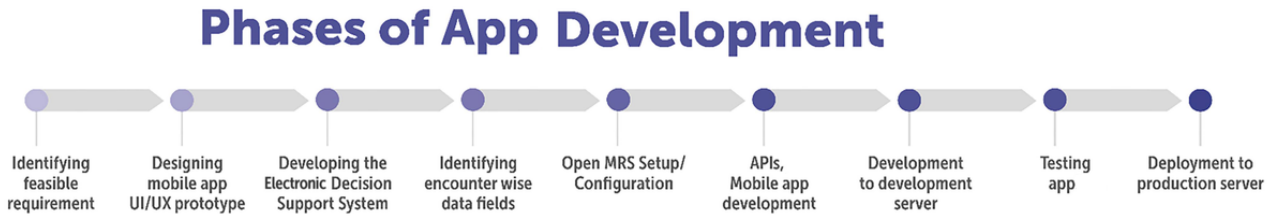
Figure 2. Workflow of data. API: application programming interface; Asha: Accredited Social Health Activist.



Phases of App Development

The apps went through different phases of development, enhancement, and adaptations across the three projects to suit the specific requirements of each project (Figure 3).

Figure 3. Phases of app development. API: application programming interface; MRS: Medical Record System; UI: user interface; UX: user experience.



Based on the scope of work (ie, a detailed document or description that outlines the specific tasks, activities, objectives, and deliverables associated with a particular project) received, the technological team assessed requirements and checked the feasibility of incorporating them.

The next step involved the design of the mobile app user interface (UI)/user experience (UX) prototype, which was an interactive mock-up of the mobile app. The prototype contained key UIs, screens, and simulated functions without any working code or final design elements. This provided a better understanding of the real-time UI and UX before production.

Subsequently, the EDSSs were designed according to standard existing diagnosis and management guidelines, which were programmed to develop the most appropriate apps. To identify encounter data fields, individual interactions by ASHAs/doctors were recorded as separate encounters in OpenMRS. Different study phases had distinct data points, necessitating a logical flow of questions. Specific roles were assigned, tailoring the data collection tools to individual responsibilities. For instance, the follow-ups for ASHAs used priority-listing questions, whereas the doctors app incorporated mhGAP tool queries. This ensured targeted and relevant data capture for each study participant.

The next step involved configuring project-specific technical details such as concepts, encounter types, visit frequencies, user roles, and API settings within OpenMRS. Additionally, custom tables were created to facilitate real-time reporting and analytics, ensuring efficient data management and analysis for the project.

The final step was the development of the mobile app and APIs, which was carried out as a multistage process. The set up followed the sequence of development, test stage, and production environments. The final prototype for the mobile app involved integrating the EDSS into the app. The SMH apps supported online/offline features. Standard security integrations were enabled while developing the mobile app in the local database in the three different environments. In the test environment, the integrated feature was assessed with test data to evaluate the impact of the load of data and the performance of the app. In the stage environment, this phase included an exact replica of a production environment for testing. In the production environment, the software or products were made live for use. Once the development of the app was complete and certified by the quality assurance team, it was deployed for the production environment. Screenshots of the app are provided in Figures 4-6.

Continuous modifications and maintenance of the app were applied across the projects' lifetimes.

Figure 4. App screenshot 1.

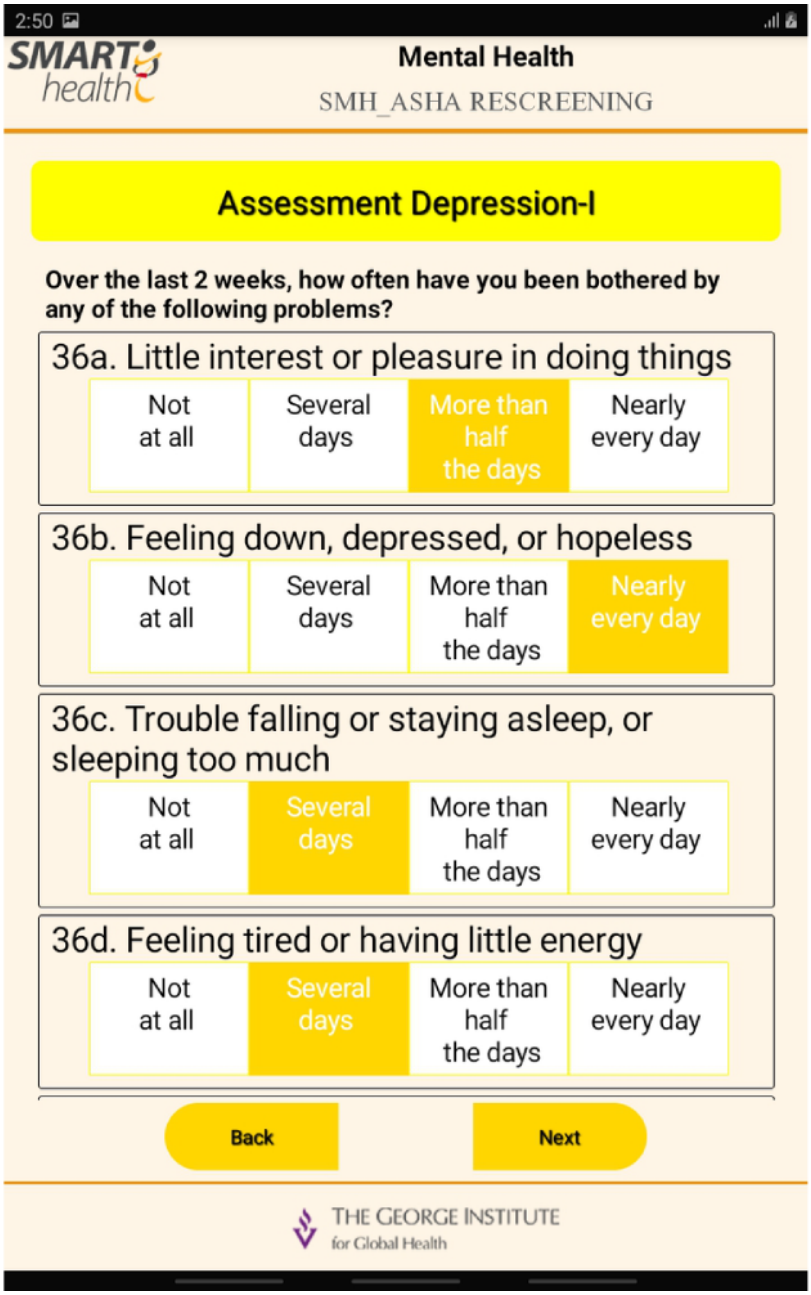


Figure 5. App screenshot 2.

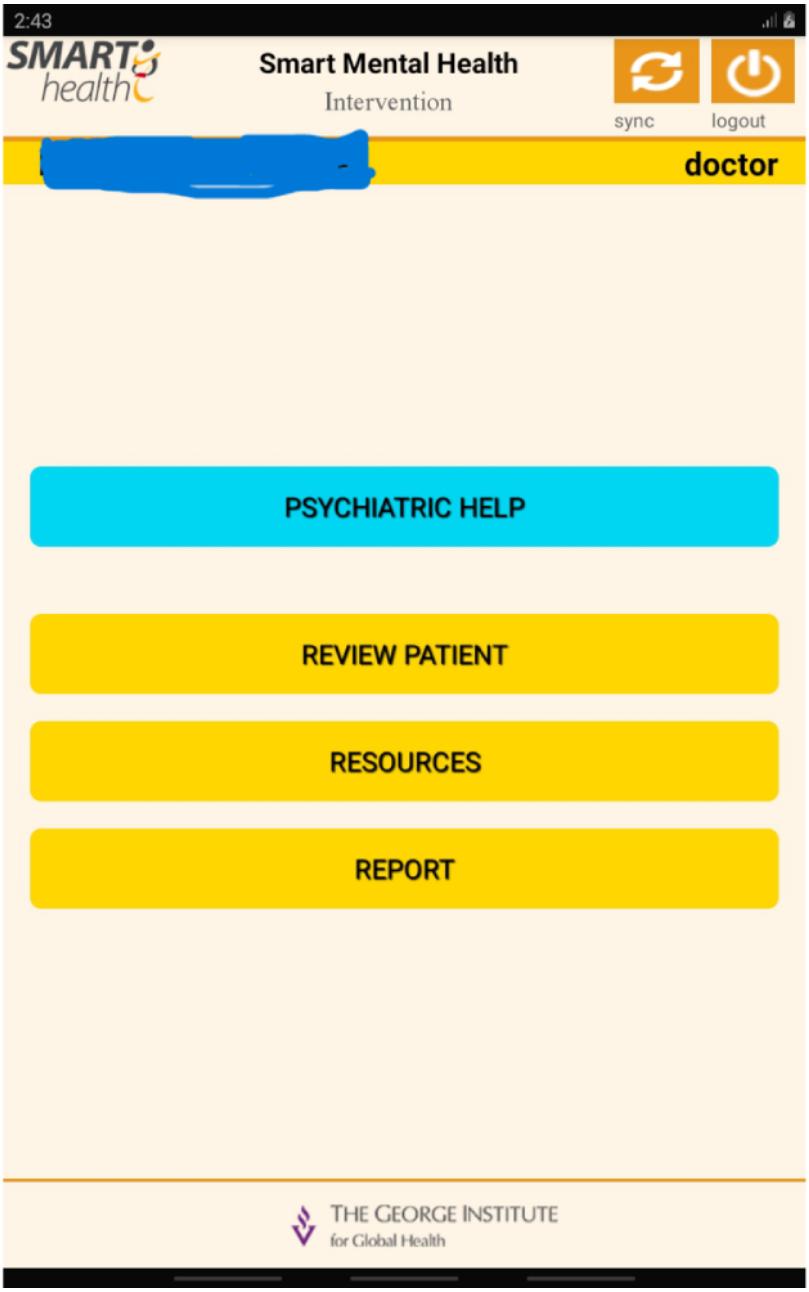
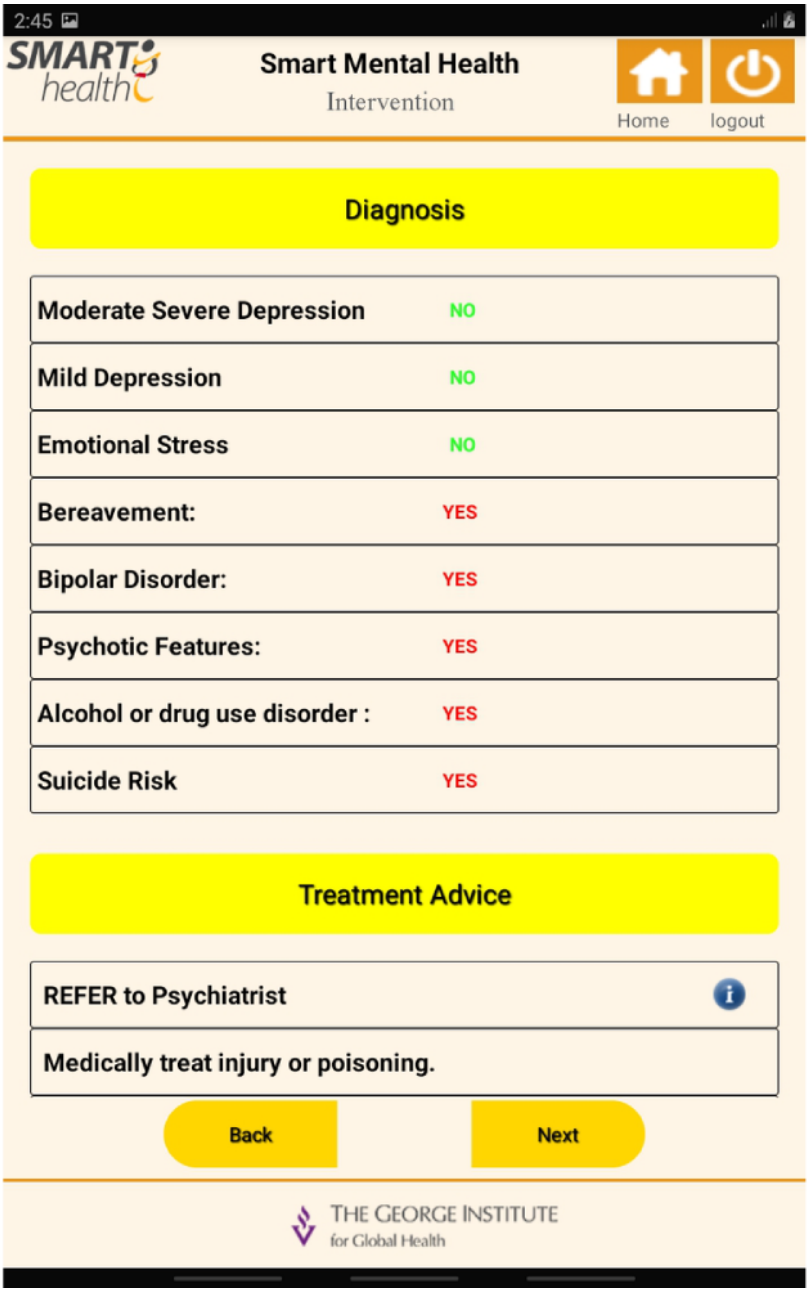


Figure 6. App screenshot 3.



Target

The SMH Pilot was implemented in 42 villages across rural and tribal areas of Andhra Pradesh [8,17] with the goal of understanding the feasibility and acceptability of using mobile technology and task-sharing approaches to address CMDs. This project covered approximately 50,000 adults and informed the subsequent SMH Trial, which took place in villages across Haryana and Andhra Pradesh, screening 165,000 adults in 133 villages and 44 PHCs. Currently, ARTEMIS is being implemented among 70,000 adolescents (10-19 years old) in 60 urban slum clusters in Vijayawada (Andhra Pradesh) and New Delhi.

Ethical Considerations

All collected data are securely stored on central servers in Hyderabad, with restricted access limited to the project team. Participants provided written consent and received detailed information about data collection at various time points. The SMART Mental Health pilot study was approved by the Independent Ethics Review Committee of the Centre for Chronic Disease Control (IRB00006330) for studies CCDC_IEC_03_2014 and CCDC_IEC_02_2014 on October 1, 2014; the SMART Mental Health cluster randomized controlled trial was approved by the George Institute Ethics Committee (009/2018) on April 27, 2018; and the ARTEMIS trial was approved by the George Institute Ethics Committee (17/2020) on September 4, 2020. The study tools were approved by The George Institute Ethics Committee, and each participant was assigned a unique identification number at the study's outset. Data were consistently deidentified before any sharing, and only research staff and the study's implementation and statistical teams had access to the data, ensuring that confidentiality and ethical standards were maintained throughout the research process.

Participating Entities

The studies have received funding from various international organizations such as Wellcome Trust/Department of Biotechnology (India Alliance), National Health and Medical Research Council Australia, and the UK Medical Research Council. Importantly, these funders are not involved in data collection or analysis and do not have access to the data. Government agencies, although collaborators, also do not manage or analyze the data. The SMH app is under intellectual property rights of the developer, The George Institute India. Local government consultation occurred for support, but they have no role in data governance.

Budget Planning

A predefined budget was allocated to the development and implementation of the technological interventions. The main costs incurred included the cost of the server (INR 500,000=US $6862) and the time cost of an Android developer and a technical lead (INR 200,000=US $2868/month for the initial 6 months for development and then a 25% time cost for maintenance). The other costs included the procurement of tablets for data collection.

Interoperability

The apps used in the three studies followed the Health Level 7 (HL7)/Fast Healthcare Interoperability Resources (FHIR) standards for exchanging patient information between a server and mobile app in JavaScript Object Notation (JSON) format. HL7 has also developed other standards, including the HL7 Clinical Document Architecture. We used FHIR in our apps as it was designed to facilitate interoperability of health care systems, allowing different health care apps and devices to easily exchange and share data. As the FHIR standard is based on modern web technologies such as Representational State Transfer principles, JSON, and Extensible Markup Language, it provides a flexible and scalable approach to health care data exchange, making it easier for developers to build interoperable apps.

Sustainability

The study was developed and implemented in collaboration with the Andhra Pradesh and Haryana governments. The tool has been previously utilized in two studies with adults while undergoing several phases of enhancements and is currently being used in the ARTEMIS study with adolescents. Poststudy, the tool will be shared with government and other nongovernmental organizations interested in using it.

Implementation (Results)

Coverage

The overall coverage of the number of study participants, ASHAs/CWVs, and doctors reached in the three studies is detailed in Table 2.

Table 2. Coverage of participants across the three projects.

Project	Study participants reached, n	ASHAs[a]/CWVs[b] included, n	Doctors included, n
SMART MH[c] Pilot (2014 to 2019)	50,000 adults	40	14
SMART MH Trial (2018 to 2022)	165,000 adults	175	50
ARTEMIS[d] (2020-2024)	69,600 adolescents (10-19 years old)	104	27

[a]ASHA: Accredited Social Health Activist.

[b]CWV: community woman volunteer.

[c]SMART MH: Systematic Medical Appraisal and Referral Treatment Mental Health.

[d]ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums.

Outcomes and Technical Amendments Made to the Data Repository

OpenMRS indicated the overall number of instances that data were collected for each individual participant at different time points in the study. As OpenMRS has a report generation model, it was difficult to compare different data points for the same person or between participants across the same time point. Hence, an intermediary database was developed in house to facilitate the process of running customized reports, which enabled comparison of data at different time points. This process evolved following considerable testing at the backend to obtain the desired output in terms of data visualization. Some customizations were made to OpenMRS to suit study requirements (Table 3).

Table 3. Steps of configurations made to the Open Medical Record System (OpenMRS).

Configuration of OpenMRS modules	Features for the study
Creation of a concept dictionary	Every data point to be used for the study was created as a concept and given a short name
Role management	The roles of each user were fixed and were restricted based on the type of activity they were expected to do; for example, the project manager was only given access to user data management and downloading reports
User management	As per our project flow, the different users were allocated to each role, such as ASHAs[a], doctors, field staff/data collector, project manager, and administrator
Encounter management	Each entry into the tab for a specific user (ie, ASHA, doctor, data collector) was recorded as an encounter with a unique encounter ID, which helped to differentiate the number of encounters that had taken place for each study participant
Managing encounter types	Based on the different phases of the study, each phase was also considered as a separate encounter, such as the screening, rescreening, ASHA follow-up, and doctor follow-up phases
Manage observations	Each data point was considered as a separate observation
Managing persons	Demographic data for every app user (ASHA, doctor) or participant were stored as person details
Managing patients	In this feature, any additional personal identifiers/demographic details identified could be modified/configured
Cohort management	Specific cohorts were created for every phase of the project, matched to the user. This enabled the users to access data of people who were in their own cohort. This helped them to identify and follow up the individuals easily. This was done both for ASHAs and doctors, with each doctor in a particular PHC[b] having a defined set of ASHAs, who in turn had a defined set of high-risk individuals
Multilocation data management	This was a custom development made to the system to ensure the data of one location (state) were not merged with data from another location. This was relevant to the SMH[c] and ARTEMIS[d] trials, which involved two different geographical locations.

[a]ASHA: Accredited Social Health Activist.

[b]PHC: primary health center.

[c]SMH: Systematic Medical Appraisal and Referral Treatment Mental Health.

[d]ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums.

In India, internet connectivity varies, particularly in rural regions. To address this, the quality of connectivity was assessed in each study area and hotspots were identified. Offline and online data storage methods were implemented, allowing local storage on tablets in areas with poor connectivity. Data could later be uploaded to the central server once connectivity was restored. Additionally, networks at certain PHCs were improved, increasing the bandwidth to enable ASHAs and doctors to upload data when in proximity to these PHCs.

Lessons Learned

There were several lessons learned while designing and implementing these interventions, which resulted in several enhancements to the systems for improving UX and achieve the study outcomes. Following the SMH Pilot, issues were identified in the EDSS that needed to be corrected for the SMH and ARTEMIS trials (Table 4). During the project, unforeseen challenges arose due to the COVID-19 pandemic. Face-to-face training for health workers was impossible, leading to the preparation of training materials delivered with the assistance of field staff. Additionally, some tablets used by health workers broke down, necessitating replacements and revealing bugs in the app. The SMH cRCT project faced difficulties because of COVID-19, and different mitigation strategies were adopted to ensure implementation of the different stages of the project. However, due to the rapidly changing situation, those also had to be modified quickly. Considering all the issues encountered earlier, we tried to mitigate all these challenges encountered during the SMH pilot study and cRCT, leading to enhancement of the apps developed for the ARTEMIS project. To have a smooth transition from the test environment to the live environment, the technical team performed additional checks by testing the apps by the field staff and creating data that were uploaded to the server to confirm whether all the fields are being populated correctly. This helped in reducing the errors while data were being captured in live scenarios.

Table 4. Enhancements made to the electronic decision support system.

Issues that needed amendments	Solutions for the problems/issues
Daily monitoring of data at the field level and comparison of data across sites, localities, and users was very difficult. Monitoring of clinical data of patients was also difficult	Development of descriptive analytics at the database level while implementing the SMH[a] trial was done to ease monitoring of data. There were many enhancements made at that level, in terms of representing real-time data from different aspects of the study. This included identification of mental health service use, the burden of different mental health conditions, and comparison of different conditions, among other factors. These analytics could be viewed by comparisons made across regions, gender, and age groups. These were represented through pictorial modes such as graphs and pie charts (see Figure 7 for examples)
Monitoring an individual's mental health status over time was not possible	Analytics were developed to track the PHQ9[b] and GAD7[c] scores of an individual in the different phases of the study. Data captured periodically during monitoring could be viewed as graphs and charts based on the longitudinal data at the backend using analytics.
The performance of ASHAs[d] could not be tracked well	There were enhancements made to the ASHA app, which tracked the performance of each ASHA and provided data about the numbers of screenings and follow-ups performed, including the time taken for each. Random audio recordings of their interactions were also captured to ensure quality checks.
As the database is encrypted and stored in a password-protected, secure location, it is hard to gain access to data by reverse engineering or decoding	The app is protected with multifactor authentication using a password and lock pin as an enhancement to the existing setup.
User interface and functioning of the app were not clear	Several changes were made to the user interface, including a change of font size, color, and creating different section headers using attractive symbols/pictures, for better user experience
Enabling online training during COVID-19	Some of the training materials were embedded in the mobile apps to enable easier access for trainees using virtual modes during COVID-19.
Real-time monitoring of the activities of field staff was required to ensure increased data quality	Random audio recording of interaction of field staff with study participants or high-risk individuals was enabled. The time taken for each screening was also made available at the database level for these audio recordings. This helped the implementation team to monitor data collection and quality.
Merging of two apps, namely household listing and participant screening, into one app	This merger made it possible for simultaneous data collection for both listing and screening, which saved time for both the participant and field staff and reduced multiple visits to the same household for data collection.

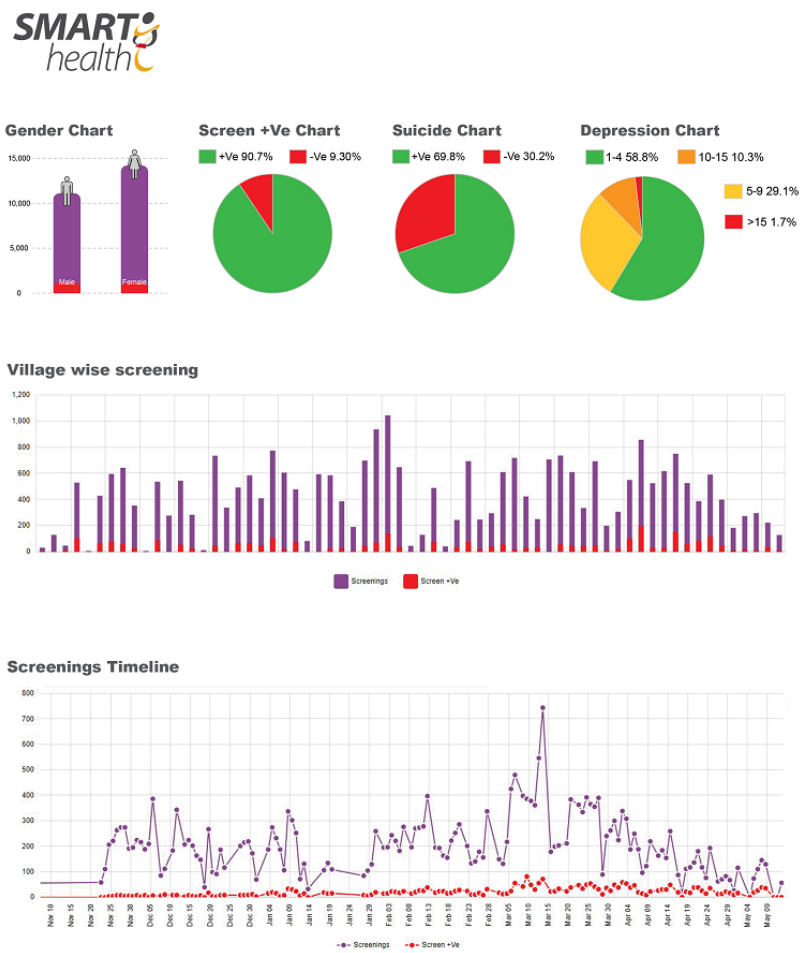
[a]SMH: Systematic Medical Appraisal and Referral Treatment Mental Health.

[b]PHQ9: Patient Health Questionnaire-9.

[c]GAD7: Generalized Anxiety Disorder-7.

[d]ASHA: Accredited Social Health Activist.

Figure 7. Snapshot of the analytics developed to monitor the progress of the study outcomes. +Ve: positive; -Ve: negative.



Discussion

Principal Findings

This paper outlines the experience of employing technology in mental health service delivery across rural and urban India in three projects. We have highlighted the implementation challenges and app adaptations based on user feedback, offering insights valuable for technology-based mental health projects in resource-limited settings. Technology-enabled interventions have shown effectiveness in diagnosing, treating, and following up on various health conditions [18,19]. Most mHealth interventions used in India have been disease-specific and do not involve a health systems approach. One example of a more health system–focused app is the Government to Government web-based monitoring information system that has been set up by the Ministry of Health & Family Welfare, Government of India, to monitor the National Health Mission and other health programs. To increase effectiveness, these innovations should focus on creating new avenues to integrate tools that have encouraging and sustainable outcomes related to access, equity, quality, and responsiveness. The SMH app can be integrated with government systems after specific modifications. The use of electronic medical record systems and telemedicine are examples of some of the interventions implemented and found to be beneficial for health care delivery for large populations, especially in LMICs [16,19,20]. However, there is a need to understand the local context and setting while developing or enhancing any existing app, as some of the original features may not be relevant to the local context, making further adaptations critical.

One way to enhance the functional capabilities of apps such as SMH is to link the app with telemedicine facilities that amplify the ability to connect to remotely located consumers with specialists located in larger cities [13]. For example, machine learning has been applied for suicide prediction, matching patients to appropriate treatment, improving the efficacy of mental health care by clinicians, and monitoring patients for treatment adherence with the help of smartphones and sensors [21].

Another way to leverage technology in mental health is by using artificial intelligence. A recent systematic review recommended the use of artificial intelligence technologies as accurate and effective strategies in the diagnosis and treatment of mental health conditions [22]. Virtual reality technology has proven to be a useful and powerful tool in addiction research [23]. The user interacts with the virtual reality environment, offering an environment close to real life that is dynamic in nature and requires active participation. These environments can be used to develop psychotherapeutic interventions by adding a personal

touch, having predictable conditions with additional features such as embodiment, eye tracking, and other biological factors [24].

There is still substantial work to be done in terms of scaling up these interventions and understanding their feasibility and acceptability across different settings and populations. Use of novel strategies such as videogaming can be explored to implement mental health interventions that can be customized to specific populations [25]. Such techniques should be considered in future iterations of the technology platform [26,27].

Limitations

There were a few limitations in our apps. First, the mobile apps developed were limited to stress, depression, anxiety, and increased suicide risk; however, the principles of including other mental health conditions would be quite similar. Second, although the projects had a system of referring participants requiring specialist care to mental health professionals, it was beyond the scope of the projects to track the care provided by the mental health professionals through our app. This was because our app was developed through primary health system–focused application for use in low-resource settings and was not linked to any central electronic health record system as is possible in more developed health systems with more robust data capture and record-sharing capabilities, such as the National Health Service in the United Kingdom or health systems in Australia. Third, the current apps are compatible on Android platforms and could not be expanded to other operating systems. Finally, the apps developed were specifically created following consultations with local stakeholders; hence, their generalizability across other settings will need to be assessed after adaptation is complete.

Future Recommendations

Given our experiences, we have compiled a set of suggested recommendations for technology-based interventions in similar settings, which are presented in Textbox 1.

Textbox 1. Recommendations for technology-based interventions.

- Inclusion of the technical team from the outset when study protocols are being developed.
- All study-related tools and database designing should be finalized in consultation with relevant experts.
- A protocol that details the process of server support in terms of setup/maintenance needs should be developed and followed.
- The server needs to factor in the size of the data set and latest versions of operating systems in reducing any issues faced.
- App user interface/user experience should be designed and assessed for acceptability by targeted populations. The use of reports or data analytics for the study must be discussed and finalized as per study needs.
- Develop systems that can be used across any kind of device, are compatible for software or version upgrades, and are web-based and easily programmable.
- A technical guide with frequently asked questions outlining the various aspects of technology, such as navigation, problem-solving, and reporting of issues, should be developed to facilitate staff training.
- The infrastructure and the architecture of the app should be flexible for making modifications or scaling up the app. The scalability is measured by the number of requests an app can manage and support the app effectively. A decision needs to be taken in terms of adding resources to the computing system for scaling either horizontally (adding more machines to the existing pool) or vertically (adding more power to the existing machines). Both types of scaling are similar as they add computing resources to the infrastructure; however, there are distinct differences between the two in terms of implementation and performance.

Conclusion

In conclusion, the development of any health-related app is subject to the context and the area where it would be implemented. There is a need for careful testing using iterative processes, allocate human and budgetary resources that are adequate, and integrate apps with larger electronic health record systems that inform health systems.

Acknowledgments

We would like to acknowledge the entire Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health team at the Andhra Pradesh, Haryana site and the staff of Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums (ARTEMIS) at the Vijayawada and Delhi sites who provided input while assessing the apps, which greatly helped in making specific enhancements to the apps. This work was supported by ARTEMIS (to SG, MG, and SK), Australian National Health and Medical Research Council (NHMRC) Global Alliances for Chronic Disease (GACD) (SMART for Common Mental Disorders in India; grant APP1143911), and UK Research and Innovation (UKRI)/Medical Research Council (MRC) grant (ARTEMIS; MR/S023224/1 to PKM). DP is partially or wholly supported through the SMART Mental Health NHMRC/GACD grant. PKM is the principal investigator on the ARTEMIS Project and coprincipal investigator on the SMART Mental Health Project and is partially supported by both projects. DP is supported by fellowships from the NHMRC of Australia (1,143,904) and the Heart Foundation of Australia (101,890). SKY is supported by the ARTEMIS Project (UKRI/MRC grant MR/S023224/1), SP is supported by the ARTEMIS Project and another project titled MentAL Health Risk Factors among Older AdolesceNts living in Urban Slums: An InTervention to Improve ResIlience (ANUMATI) funded by the Indian Council of Medical Research

(grant 2019-0531). MD is supported by SMART Mental Health funded by NHMRC Australia (grant APP1143911) and the International Study of Discrimination and Stigma Outcomes (INDIGO) Partnership Research Programme funded by the UK MRC (MR/R023697/1). The funding bodies played no role in the design of the study and in the conceptualization and writing of the manuscript.

Authors' Contributions

SK, SG, PKM: conceptualization. SK and SG: writing of first draft. MG, SK, SKY, SP, MD, DP, and PM: review & editing. All authors read and approved the final manuscript.

Conflicts of Interest

All authors are employees of The George Institute, which has a part-owned social enterprise, George Health Enterprises, with commercial relationships involving digital health innovations.

References

1. Vigo D, Thornicroft G, Atun R. Estimating the true global burden of mental illness. Lancet Psychiatry 2016 Feb;3(2):171-178. [doi: 10.1016/S2215-0366(15)00505-2] [Medline: 26851330]
2. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bull World Health Organ 2004 Nov;82(11):858-866 [FREE Full text] [Medline: 15640922]
3. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry 2017 Feb 02;210(2):119-124 [FREE Full text] [doi: 10.1192/bjp.bp.116.188078] [Medline: 27908899]
4. Naslund JA, Marsch LA, McHugo GJ, Bartels SJ. Emerging mHealth and eHealth interventions for serious mental illness: a review of the literature. J Ment Health 2015 May 28;24(5):321-332 [FREE Full text] [doi: 10.3109/09638237.2015.1019054] [Medline: 26017625]
5. World Health Organization. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. Geneva: World Health Organization; 2011.
6. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236 [FREE Full text] [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
7. Agrawal N. Telephone network and internet penetration in India: a pragmatic study using data analytics. Global J Enterprise Inf Syst 2021;13(1):42-48 [FREE Full text]
8. Maulik PK, Devarapalli S, Kallakuri S, Bhattacharya A, Peiris D, Patel A. The systematic medical appraisal referral and treatment mental health project: quasi-experimental study to evaluate a technology-enabled mental health services delivery model implemented in rural India. J Med Internet Res 2020 Feb 27;22(2):e15553 [FREE Full text] [doi: 10.2196/15553] [Medline: 32130125]
9. Daniel M, Maulik PK, Kallakuri S, Kaur A, Devarapalli S, Mukherjee A, et al. An integrated community and primary healthcare worker intervention to reduce stigma and improve management of common mental disorders in rural India: protocol for the SMART Mental Health programme. Trials 2021 Mar 02;22(1):179 [FREE Full text] [doi: 10.1186/s13063-021-05136-5] [Medline: 33653406]
10. Yatirajula SK, Kallakuri S, Paslawar S, Mukherjee A, Bhattacharya A, Chatterjee S, et al. An intervention to reduce stigma and improve management of depression, risk of suicide/self-harm and other significant emotional or medically unexplained complaints among adolescents living in urban slums: protocol for the ARTEMIS project. Trials 2022 Jul 29;23(1):612 [FREE Full text] [doi: 10.1186/s13063-022-06539-8] [Medline: 35906663]
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:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
12. Maulik PK, Tewari A, Devarapalli S, Kallakuri S, Patel A. The Systematic Medical Appraisal, Referral and Treatment (SMART) Mental Health Project: development and testing of electronic decision support system and formative research to understand perceptions about mental health in rural India. PLoS One 2016 Oct 12;11(10):e0164404 [FREE Full text] [doi: 10.1371/journal.pone.0164404] [Medline: 27732652]
13. Daniel M, Kaur A, Mukherjee A, Bhattacharya A, Tewari A, Sagar R, et al. The systematic medical appraisal, referral and treatment (SMART) mental health programme: formative research informing a cluster randomized controlled trial. SSM Mental Health 2023 Dec;3:100223. [doi: 10.1016/j.ssmmh.2023.100223]
14. Scaling up care for mental, neurological and substance use disorders: mhGAP. World Health Organization. 2008. URL: https://www.who.int/activities/scaling-up-mental-health-care [accessed 2024-01-14]
15. OpenMRS. URL: https://openmrs.org/ [accessed 2024-01-22]
16. Peiris D, Praveen D, Mogulluru K, Ameer MA, Raghu A, Li Q, et al. SMARThealth India: a stepped-wedge, cluster randomised controlled trial of a community health worker managed mobile health intervention for people assessed at high

cardiovascular disease risk in rural India. PLoS One 2019 Mar 26;14(3):e0213708 [FREE Full text] [doi: 10.1371/journal.pone.0213708] [Medline: 30913216]
17. Maulik PK, Kallakuri S, Devarapalli S, Vadlamani VK, Jha V, Patel A. Increasing use of mental health services in remote areas using mobile technology: a pre-post evaluation of the SMART Mental Health project in rural India. J Glob Health 2017 Jun;7(1):010408 [FREE Full text] [doi: 10.7189/jogh.07.010408] [Medline: 28400954]
18. Bassi A, John O, Praveen D, Maulik PK, Panda R, Jha V. Current status and future directions of mHealth interventions for health system strengthening in India: systematic review. JMIR Mhealth Uhealth 2018 Oct 26;6(10):e11440 [FREE Full text] [doi: 10.2196/11440] [Medline: 30368435]
19. Koppar AR, Sridhar V. A workflow solution for electronic health records to improve healthcare delivery efficiency in rural India. 2009 Presented at: 2009 International Conference on eHealth, Telemedicine, and Social Medicine; February 1-7, 2009; Cancun, Mexico. [doi: 10.1109/etelemed.2009.30]
20. Acharya R, Rai J. Evaluation of patient and doctor perception toward the use of telemedicine in Apollo Tele Health Services, India. J Family Med Prim Care 2016;5(4):798-803 [FREE Full text] [doi: 10.4103/2249-4863.201174] [Medline: 28348994]
21. Haggerty E. Healthcare and digital transformation. Network Security 2017 Aug;2017(8):7-11. [doi: 10.1016/s1353-4858(17)30081-8]
22. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim H, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019 Nov 07;21(11):116 [FREE Full text] [doi: 10.1007/s11920-019-1094-0] [Medline: 31701320]
23. Segawa T, Baudry T, Bourla A, Blanc J, Peretti C, Mouchabac S, et al. Virtual reality (VR) in assessment and treatment of addictive disorders: a systematic review. Front Neurosci 2019 Jan 10;13:1409 [FREE Full text] [doi: 10.3389/fnins.2019.01409] [Medline: 31998066]
24. Cavalcante Passos I, Mwangi B, Kapczinski F. Personalized psychiatry: Big data analytics in mental health. New York, NY: SpringerLink; 2019.
25. Hamari J, Keronen L. Why do people play games? A meta-analysis. Int J Inf Manag 2017 Jun;37(3):125-141. [doi: 10.1016/j.ijinfomgt.2017.01.006]
26. Wilkinson N, Ang RP, Goh DH. Online video game therapy for mental health concerns: a review. Int J Soc Psychiatry 2008 Jul 01;54(4):370-382. [doi: 10.1177/0020764008091659] [Medline: 18720897]
27. Li J, Theng Y, Foo S. Game-based digital interventions for depression therapy: a systematic review and meta-analysis. Cyberpsychol Behav Soc Netw 2014 Aug;17(8):519-527 [FREE Full text] [doi: 10.1089/cyber.2013.0481] [Medline: 24810933]

Abbreviations

API: application programming interface
ARTEMIS: Adolescents' Resilience and Treatment Needs for Mental Health in Indian Slums
ASHA: Accredited Social Health Activist
CMD: common mental disorder
CWV: community woman volunteer
cRCT: cluster randomized controlled trial
EDSS: electronic decision support system
FHIR: Fast Healthcare Interoperability Resource
HL7: Health Level 7
JSON: JavaScript Object Notation
LMIC: low- and middle-income country
mHealth: mobile health
OpenMRS: Open Medical Record System
PHC: primary health center
SMH: Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health
UI: user interface
UX: user experience
WHO: World Health Organization

Edited by C Perrin; submitted 22.03.23; peer-reviewed by N Mungoli, E Korshakova; comments to author 24.04.23; revised version received 21.05.23; accepted 29.11.23; published 15.02.24.

Please cite as:
Kallakuri S, Gara S, Godi M, Yatirajula SK, Paslawar S, Daniel M, Peiris D, Maulik PK
Learnings From Implementation of Technology-Enabled Mental Health Interventions in India: Implementation Report
JMIR Med Inform 2024;12:e47504
URL: https://medinform.jmir.org/2024/1/e47504
doi:10.2196/47504
PMID:38358790

Implementation Report

A Mobile App (Concerto) to Empower Hospitalized Patients in a Swiss University Hospital: Development, Design, and Implementation Report

Damien Dietrich[1,2], MD; Helena Bornet dit Vorgeat[3], BSc, MBA; Caroline Perrin Franck[1], PhD; Quentin Ligier[3], MSc

[1]Geneva Hub for Global Digital Health, Faculty of Medicine, University of Geneva, Geneva, Switzerland

[2]Kheops Technologies SA, Plan-Les-Ouates, Switzerland

[3]Geneva University Hospitals, Geneva, Switzerland

Corresponding Author:
Damien Dietrich, MD
Geneva Hub for Global Digital Health
Faculty of Medicine
University of Geneva
Campus Biotech
9 Chemin des Mines
Geneva, 1202
Switzerland
Phone: 41 227714730
Email: damien.dietrich@gmail.com

Abstract

Background: Patient empowerment can be associated with better health outcomes, especially in the management of chronic diseases. Digital health has the potential to promote patient empowerment.

Objective: Concerto is a mobile app designed to promote patient empowerment in an in-patient setting. This implementation report focuses on the lessons learned during its implementation.

Methods: The app was conceptualized and prototyped during a hackathon. Concerto uses hospital information system (HIS) data to offer the following key functionalities: a care schedule, targeted medical information, practical information, information about the on-duty care team, and a medical round preparation module. Funding was obtained following a feasibility study, and the app was developed and implemented in four pilot divisions of a Swiss University Hospital using institution-owned tablets.

Implementation (Results): The project lasted for 2 years with effective implementation in the four pilot divisions and was maintained within budget. The induced workload on caregivers impaired project sustainability and warranted a change in our implementation strategy. The presence of a killer function would have facilitated the deployment. Furthermore, our experience is in line with the well-accepted need for both high-quality user training and a suitable selection of superusers. Finally, by presenting HIS data directly to the patient, Concerto highlighted the data that are not fit for purpose and triggered data curation and standardization initiatives.

Conclusions: This implementation report presents a real-world example of designing, developing, and implementing a patient-empowering mobile app in a university hospital in-patient setting with a particular focus on the lessons learned. One limitation of the study is the lack of definition of a “key success” indicator.

(JMIR Med Inform 2024;12:e47914) doi:10.2196/47914

KEYWORDS

patient empowerment; mobile apps; digital health; mobile health; implementation science; health care system; hospital information system; health promotion

Introduction

Context

During recent decades, medicine has been moving from a focus on paternalistic approaches toward a paradigm of patient-centeredness, highlighting patient partnership and participation. Patient empowerment refers to a metaconcept with no unique definition [1]. However, it is commonly accepted that empowered patients possess key capacities and resources to be able to (1) participate in shared decision-making, (2) manage their own health, and (3) self-empower themselves [1].

Patient Empowerment and Clinical Outcomes

Some studies have demonstrated a positive association between patient empowerment and improved clinical outcomes or their proxy. This is best documented in the context of chronic diseases, especially diabetes. Wong et al [2] compared serum glycated hemoglobin (HbA_{1C}) and low-density lipoprotein cholesterol (LDL-C) levels in a group following implementation of a patient empowerment program (PEP) or the standard of care, resulting in decreased LDL-C levels in the PEP group. Similarly, Lian et al [3] found a lower incidence of all-cause mortality, cardiovascular events, and diabetes mellitus complications following participation in a PEP. In a review of randomized controlled trials, a decrease in HbA_{1C} and blood pressure levels was associated with empowerment interventions for patients with diabetes in sub-Saharan Africa [4]. In a meta-analysis, Baldoni et al [5] reported an improvement in HbA_{1C} levels following collective empowerment strategies. In a systematic review, Shnaigat et al [6] identified patient activation, a concept related to empowerment, as a valid strategy to improve outcomes of patients with chronic obstructive pulmonary disease.

However, it is important to highlight that several studies also reported no beneficial effects of empowerment programs. A 2017 meta-analysis found no statistically significant positive effect of empowerment on HbA_{1C} levels, despite five included studies reporting positive results [7]. Santoso et al [8] reported a lack of evidence to demonstrate a positive association between women's empowerment and outcomes of child nutrition.

The lack of clear definitions and measures for empowerment may explain these controversial findings. Differences in program design could also contribute to this variability; therefore, further research identifying determinants for a successful intervention is needed. Indeed, the authors of the cited studies often reported the poor availability of high-quality research.

Digital Health and Patient Empowerment

With the variety of solutions that could be envisioned, digital health is seen as a promising tool to promote patient empowerment and, indirectly, outcomes. However, mixed results are seen in the related literature.

In a systematic review, Johansson et al [9] showed that online communities support patient empowerment. Sosa et al [10] reported that a text messaging–based empowering intervention following head and neck surgery was both highly appreciated by patients and feasible. Conversely, Ammenwerth et al [11] reported no clinically relevant effect of patient portals on patient empowerment or health-related outcomes in a systematic review. Vitger et al [12] failed to demonstrate a positive effect of digital interventions to support shared decision-making, which was likely due to the small number of high-quality studies available. Verweel et al [13] found limited evidence demonstrating a positive effect of a digital intervention for health literacy. Finally, Thomas et al [14] reported that the quality and adequacy of the content of patient-empowering mobile apps varied greatly, urging for a more rigorous design and further testing before implementation. To our knowledge, no study has directly shown a link between mobile health app–induced empowerment and direct health outcomes.

Overall, few high-quality studies assessing the effect of digital health interventions on patient empowerment are available. Research is needed to confirm or deny the high perceived potential of digital tools.

Concerto: A Mobile App Designed to Promote Patient Empowerment

Concerto is a mobile app designed to promote the empowerment of hospitalized patients. The app was initially designed during a hackathon in 2015 by a multidisciplinary team including health care and IT professionals as well as one patient. Building on the hackathon prototype and after a feasibility study, the Geneva University Hospitals (HUG) launched a project aiming at developing and implementing a fully functional mobile app delivered on institution-owned tablets in four pilot divisions (oncology, neurorehabilitation, orthopedics, and pediatrics) and assessing its effectiveness. Following this pilot study, the mobile app was further refined and deployed institution-wide based on a bring-your-own-device (BYOD) approach. This implementation report focuses on the pilot study only, with the objective to highlight the lessons learned. The report is structured following the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) guidelines [15].

Methods

Design and Agile Development

Building on the prototype developed during the hackathon, the foreseen functionalities of Concerto were first compared with patients' expectations using focus groups and a semiquantitative questionnaire. A feasibility study was then performed to assess the availability and quality of the necessary data in the hospital health information system (HIS), which has been developed mainly in-house during the last 30 years.

Based on the patients' insights, further described in Dietrich et al [16], version 1.0 of Concerto was specified and developed using an agile methodology with frequent user testing among hospitalized patients. The main functionalities of this version of Concerto included:

1. An up-to-date calendar on which patients can visualize their care schedule and better understand their daily planning with the aim to reduce the impact of these events and be better prepared for them.

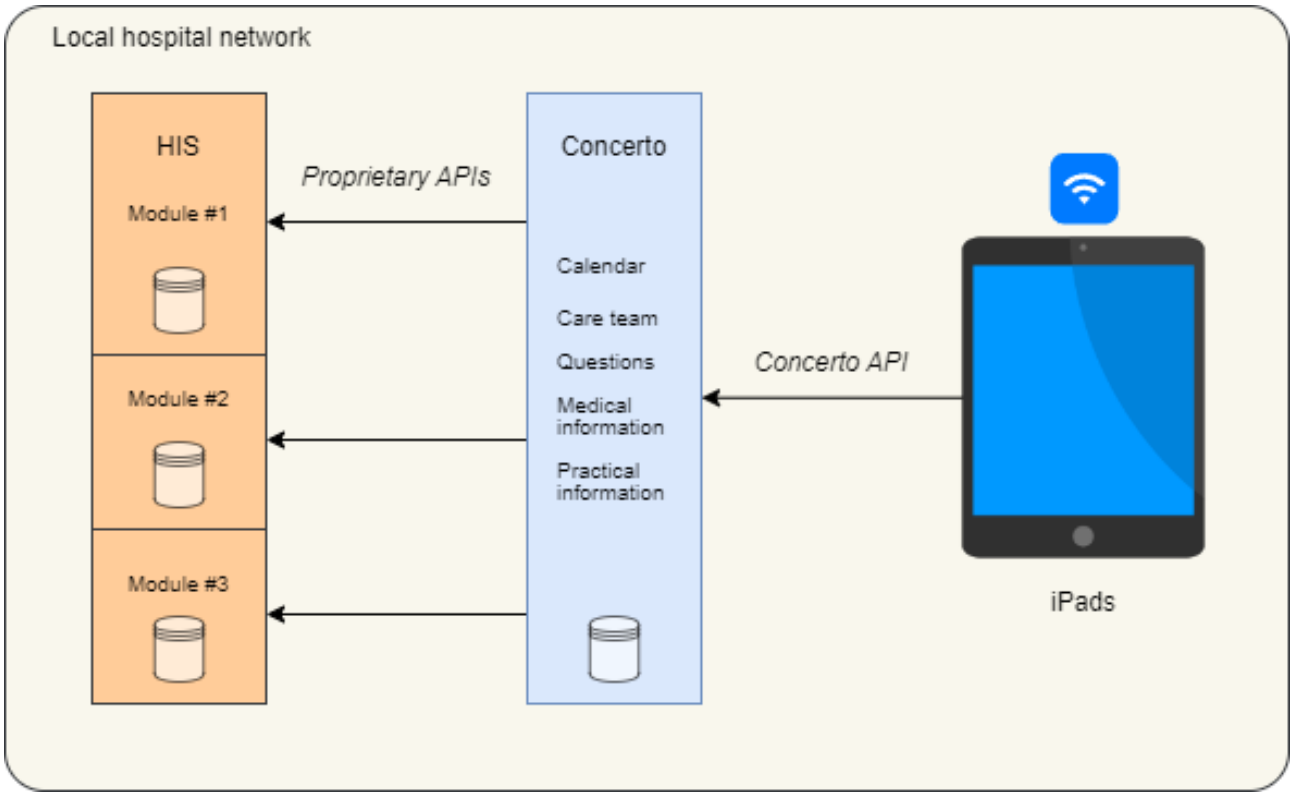
2. A care team module on which the patient can obtain information about their on-duty care team, including names and photographs, to better know the professionals they will meet during their stay and facilitate communication.
3. A “questions” module on which patients can prepare their questions for health care professionals during the medical round and thus elevate the level of communication.
4. An “information” module on which patients have access to targeted medical information. Expected benefits of this module were to achieve better situation awareness, better treatment adherence, and early detection of complications.
5. A “practical information” module on which patients can find useful information about their stay at the hospital to improve their overall experience.
6. A social network module on which patients can interact with HUG accounts.

In subsequent versions, a new module was added, allowing the patient to choose their meal directly on the app rather than via a form completed by the nurse. This module was designed to simultaneously improve the patient experience while decreasing the nurse workload.

The app was developed in web languages (an Angular project), encapsulated as an iOS app (with Apache Cordova), and deployed on institution-owned tablets using a mobile device management solution. The key arguments for internal development over acquisition of a commercial solution were that (1) a significant part of the development work was about interfacing with the HIS, and (2) to our knowledge, no adequate and mature commercial solution was available at the time, although such solutions have emerged since then.

The initial version of the app connected directly to the custom-made HUG HIS using its proprietary interfaces for the sake of development simplicity and to alleviate time constraints. Further versions of the app have used industry standards such as Health Level 7 Fast Healthcare Interoperability Resources, with the vision to enable Concerto to connect more easily to other HISs in the future. This update has required new developments on the HIS side and was not achievable during the pilot phase described in this report. Figure 1 presents the simplified architecture of Concerto.

Figure 1. Technical overview of Concerto. Through the hospital's private Wi-Fi, the patient's tablet connects to the Concerto server, which contains some patient-generated data and the business logic to provide the app data. The server connects to different modules of the hospital information system (HIS) to retrieve other data using diverse application programming interfaces (APIs).



Implementation

The definition of the logistics necessary to deliver Concerto on institution-managed tablets was an important part of the project. The following process was repeated for each patient: (1) setting up the tablet, including defining a personal passcode; (2) two-factor authentication in the Concerto app using the patient ID, scanned from the identification bracelet, and an SMS text messaging challenge; (3) on-demand charging; (4) disinfecting the tablet after the patient’s discharge; and (5) reinitializing and erasing the tablet. Tablets were charged and stored under key-secured storage in the nurse office. Each tablet was protected using individual cases. Hygiene procedures were validated by the Infection Prevention and Control Division of the HUG.

Once version 1.0 became available, caregivers of the different divisions were trained for 30 minutes during hands-on sessions

in which (1) the project and app were presented; (2) the logistics of the tablets were explained; and (3) most importantly, they had the opportunity to familiarize themselves with the tool. At least one caregiver was defined as a "superuser" on a voluntary basis and was implicated from the beginning of the project. The specific responsibilities of superusers included (1) acquiring deep knowledge of the app, (2) being the focal point for exchange with the project team, and (3) acting as the referent for day-to-day questions of caregivers. A typical division included 20 beds and comprised a pool of over 50 caregivers that were trained during different sessions. Importantly, as in many hospital projects, caregivers did not have dedicated time for the project. Therefore, they had to manage making themselves available during a normal day of work.

One unit was scheduled for launch every 2 weeks, with constant presence of one member of the Concerto team during the first few days. Only patients able to interact with a mobile app, as assessed by their caregivers, were offered to use the app. To this end, caregivers used a communication flyer describing the functionalities of the app, the modalities of its use, as well as data and privacy considerations.

Bugs, feedback, and general satisfaction were systematically consigned to fuel the improvement-and-fix backlog.

Data Considerations

At the stage presented during preparation of this report, Concerto worked mainly in "read-only" mode for personal health data available in the HIS and for insensitive, impersonal information. The information patients accessed from the HIS was part of their medical records. According to Swiss law, every patient owns the data contained in their medical record, except for personal notes of health care professionals, which were out of the scope of Concerto. Accordingly, Concerto facilitated access to data already owned by the patients.

The access to this sensitive personal information required a secure log-in based on the patient's ID number and a second-factor authentication with an SMS text challenge. The use of institution-owned devices allowed Concerto to access data in the hospital's local network, preventing unwanted access from the rest of the world.

The only personal information entered in Concerto included any questions patients may have had before interacting with their caregivers. This information was stored in the HIS and deleted after the hospital stay. Tablets were erased and reinitialized between patients, ensuring that no information leakage was possible between patients using the same tablet.

To summarize, Concerto facilitated the access to personal health information owned by the patient without the possibility to modify information from the app, and further allowed the patient to enter personal health information stored in the HIS that is inaccessible to others with all information systematically erased after the patient's hospital stay.

Overall, the project was compliant with the Swiss Law for Data Protection [17].

Funding and Budget Planning

The feasibility study and initial concept were self-funded by the eHealth and Telemedicine Division of the HUG, with the budget including salaries for a junior developer and a senior project manager.

The pilot project was then funded by private foundations based in Geneva, which included the salaries as well as necessary materials (tablets, covers, and software licenses).

Overall, the order of magnitude of the project costs ranged between US $150,000 and US $200,000, from which 25% was used for materials.

Ethical Considerations

This study is based on an internal project of a Swiss University Hospital, aiming for quality improvement. As such, no patient or participant was included specifically for this study. Moreover, no patient data of any kind were collected. Accordingly, this study does not qualify for a review by the Geneva Canton Ethics Board (Commission Cantonale d'Éthique de la Recherche sur l'Être Humain [18]). As there were no participants involved in the research, no consent, compensation scheme, or privacy and confidentiality considerations applies.

Implementation (Results)

Project Summary

Concerto was implemented in four pilot divisions; a typical division includes 20 beds and comprises a pool of over 50 caregivers.

The timeline of the various stages of the project is provided in Figure 2. From the initial hackathon to acquiring the funding, approximately 1 year was necessary to refine the concept with patients and assess the feasibility of the app. Following funding acquisition, 6 months of development were needed, followed by 6 months of piloting in the four selected hospital divisions. Overall, the project took 2 years.

The budget was respected. However, additional funding would have been welcome to help free the caregivers from their clinical duties to enable better implementation (see below for further discussion of this point).

The development team considered the agile development phase to be efficient and productive.

Critical to the development process, the organization of focus groups and one-on-one interviews with patients were facilitated owing to the clinical background of the project manager. The development team reported that early contacts with the IT division during the feasibility study helped to improve communication and hence efficiency. Finally, dedicated support of the management unlocked political stalling.

During the pilot phase, the app was proposed to all eligible patients (see the Implementation subsection of the Methods). The percentage of eligible patients was unfortunately not systematically monitored but varied according to the profile of hospitalized patients across the different divisions and over

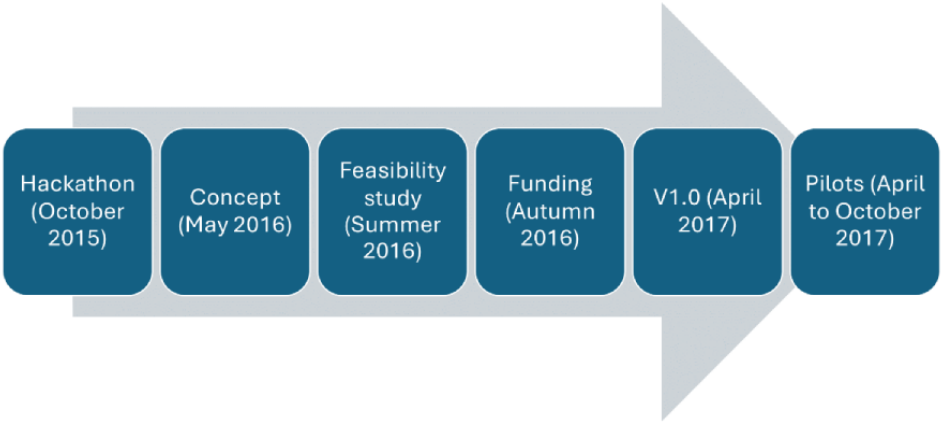
time. An eligibility rate below 50% of all hospitalized patients was common.

The percentage of acceptance was also not systematically collected. However, the project team recalled acceptance to be relatively less variable than the percentage of eligible patients and consistently high (over 80%).

The dropout rate should have also been monitored carefully to identify the reasons for dropping out.

Most importantly, navigating the logistics of the tablet emerged as a particular challenge for caregivers. Despite the support of the project team, this impaired the inclusion of patients and consequently use of the app. More precisely, caregivers reported difficulties in assisting patients with the log-in and reinitialization procedures, and all logistics steps were reported as being too time-consuming. Based on this finding, it was decided to stop the pilot phase and transition to a BYOD approach.

Figure 2. Timeline of the main phases of the project.



Lessons Learned and Determinants of Success and Failure

As often reported in the field, most challenges were encountered during the implementation (pilot) phase.

Our strategy to use institution-owned tablets added an important workload on the care teams because they were in charge of managing the tablet fleets in their divisions. This strategy was based on a rationale for cybersecurity and development; however, we underestimated the additional work it would generate for already overwhelmed caregivers. With such a strategy, it is our experience that protected and dedicated time for training caregivers is mandatory, at least for superusers. It is well recognized that the quality of training represents a key success factor for the implementation of electronic medical records (EMRs) [19]. Our impression is that this also applies to our project. Accordingly, our two first reported determinants of success are (1) having an implementation strategy that minimizes the impact on already overworked health care professionals and (2) including quality training time protected from the daily routine.

Similarly, we noticed that implementation was easier in care units in which the superuser was both convinced of the project's benefits and was an influential figure among their peers. Accordingly, our third observed determinant of success is that the presence of a "killer function," which on its own brings tremendous value, would have increased adoption by stakeholders. Even though such a function was not identified during patient focus groups, it was revealed during the implementation as the possibility for the patient to choose and order their own meals. Indeed, this function had the potential to both empower the patient and free up time for the caregivers. We consider that having such a functionality will be particularly relevant before the full-scale implementation.

The communication with the project's stakeholders was considered to be a key factor to maintain motivation and trust in the project. In particular, reactivity in fixing identified bugs or transparency about delays was appreciated.

Finally, we realized that the quality of the HIS medical information fueling Concerto was not always appropriate for display in a patient mobile app. This was either because the information was not timely or was incorrect in some cases, but most importantly because its label was too technical. This issue was associated with disadvantages such as a lack of confidence in the project as well as advantages such as a welcome transparency about HIS data, triggering continuous improvements. For example, specific agenda labels designed for patients were created in the HIS owing to the implementation of Concerto.

Discussion

This implementation report presents a real-world example of designing, developing, and implementing a patient-empowering mobile app in an in-patient setting of a Swiss public university hospital. The lessons learned, as presented in the Implementation (Results) section, are summarized in Table 1.

As described in the Introduction section, patient empowerment is a metaconcept. Hence, it is difficult to monitor with a single indicator. For this reason, a key success indicator was not defined at the beginning of this project, which has complicated its evaluation. This represents a limitation of this report, as an objective metric would have been important for complete evaluation. Simple monitoring metrics (eg, eligibility, number of users, and dropout and acceptance ratios) should have also

been collected and are planned for the next app version. A randomized controlled trial assessing the effectiveness of the Concerto mobile app on a patient situation awareness score has been designed and should be conducted in the near future. This trial will allow for better evaluation of the cost-effectiveness of such a project. Overall, data on the effectiveness of eHealth projects are often lacking, and the creation of a dedicated "Implementation Report" article type in *JMIR Medical Informatics* is helping to fill this gap.

The generalizability of our study is another limitation. Indeed, the innovation ecosystem and the EMR landscape at the HUG are very specific and different constraints may be experienced in other settings. However, we believe the reported lessons learned remain relevant in various environments.

In response to one of the main lessons learned with the pilot implementation of Concerto, a BYOD version of the app was developed. With this version, every patient was able to use the app on their personal devices, including computers, tablets, or smartphones. This decision was made to limit the workload on caregivers and improve the adoption rate. New functionalities such as the possibility for patients to choose their meal were also developed to answer unmet needs for both end users and stakeholders impacted by implementation of the app (ie, caregivers). Important challenges in terms of cybersecurity, interoperability, and compatibility had to be met with development of the BYOD version. These will be further described in a forthcoming implementation report focusing on this project phase.

Table 1. Main lessons learned and associated perceived relevance.

Lessons learned	Perceived relevance[a]
Minimize the workload of caregivers or, if possible, decrease it	5/5
Plan protected time for training end users	4/5
Select a convinced and influential superuser	3/5
Wait for a killer function before implementing the app	5/5
Maintain trust through reactivity and transparent communication	4/5

[a]Based on perceived experience, lessons learned were identified by the authors and their relevance was assessed by consensus using a score ranging from 1 (minimally important) to 5 (maximally important).

Acknowledgments

The Fondation Privée des Hôpitaux Universitaires de Genève was the main sponsor for the development and implementation of Concerto as described in this paper.

Authors' Contributions

DD was the project manager for Concerto during the project phases described in this report and wrote the manuscript. HBdV has been the project manager for Concerto after the project phases described in this implementation report and reviewed the manuscript. CPF has reviewed the manuscript. QL has been the lead developer of Concerto during the project phases described in this report and reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Bravo P, Edwards A, Barr PJ, Scholl I, Elwyn G, McAllister M, Cochrane Healthcare Quality Research Group, Cardiff University. Conceptualising patient empowerment: a mixed methods study. BMC Health Serv Res 2015 Jul 01;15:252 [FREE Full text] [doi: 10.1186/s12913-015-0907-z] [Medline: 26126998]
2. Wong CKH, Wong WCW, Lam CLK, Wan YF, Wong WHT, Chung KL, et al. Effects of patient empowerment programme (PEP) on clinical outcomes and health service utilization in type 2 diabetes mellitus in primary care: an observational matched cohort study. PLoS One 2014 May 1;9(5):e95328 [FREE Full text] [doi: 10.1371/journal.pone.0095328] [Medline: 24788804]
3. Lian J, McGhee SM, So C, Chau J, Wong CKH, Wong WCW, et al. Five-year cost-effectiveness of the Patient Empowerment Programme (PEP) for type 2 diabetes mellitus in primary care. Diabetes Obes Metab 2017 Sep 05;19(9):1312-1316. [doi: 10.1111/dom.12919] [Medline: 28230312]
4. Mogueo A, Oga-Omenka C, Hatem M, Kuate Defo B. Effectiveness of interventions based on patient empowerment in the control of type 2 diabetes in sub-Saharan Africa: A review of randomized controlled trials. Endocrinol Diabetes Metab 2021 Jan 25;4(1):e00174 [FREE Full text] [doi: 10.1002/edm2.174] [Medline: 33532614]

5. Baldoni NR, Aquino JA, Sanches-Giraud C, Di Lorenzo Oliveira C, de Figueiredo RC, Cardoso CS, et al. Collective empowerment strategies for patients with diabetes mellitus: a systematic review and meta-analysis. Prim Care Diabetes 2017 Apr;11(2):201-211. [doi: 10.1016/j.pcd.2016.09.006] [Medline: 27780683]
6. Shnaigat M, Downie S, Hosseinzadeh H. Effectiveness of patient activation interventions on chronic obstructive pulmonary disease self-management outcomes: a systematic review. Aust J Rural Health 2022 Feb 16;30(1):8-21. [doi: 10.1111/ajr.12828] [Medline: 35034409]
7. Aquino JA, Baldoni NR, Flôr CR, Sanches C, Di Lorenzo Oliveira C, Alves GCS, et al. Effectiveness of individual strategies for the empowerment of patients with diabetes mellitus: a systematic review with meta-analysis. Prim Care Diabetes 2018 Apr;12(2):97-110. [doi: 10.1016/j.pcd.2017.10.004] [Medline: 29162491]
8. Santoso MV, Kerr RB, Hoddinott J, Garigipati P, Olmos S, Young SL. Role of women's empowerment in child nutrition outcomes: a systematic review. Adv Nutr 2019 Nov 01;10(6):1138-1151 [FREE Full text] [doi: 10.1093/advances/nmz056] [Medline: 31298299]
9. Johansson V, Islind AS, Lindroth T, Angenete E, Gellerstedt M. Online communities as a driver for patient empowerment: systematic review. J Med Internet Res 2021 Feb 09;23(2):e19910 [FREE Full text] [doi: 10.2196/19910] [Medline: 33560233]
10. Sosa A, Heineman N, Thomas K, Tang K, Feinstein M, Martin MY, et al. Improving patient health engagement with mobile texting: a pilot study in the head and neck postoperative setting. Head Neck 2017 May 06;39(5):988-995 [FREE Full text] [doi: 10.1002/hed.24718] [Medline: 28263468]
11. Ammenwerth E, Hoerbst A, Lannig S, Mueller G, Siebert U, Schnell-Inderst P. Effects of adult patient portals on patient empowerment and health-related outcomes: a systematic review. Stud Health Technol Inform 2019 Aug 21;264:1106-1110. [doi: 10.3233/SHTI190397] [Medline: 31438096]
12. Vitger T, Korsbek L, Austin SF, Petersen L, Nordentoft M, Hjorthøj C. Digital shared decision-making interventions in mental healthcare: a systematic review and meta-analysis. Front Psychiatry 2021 Sep 6;12:691251 [FREE Full text] [doi: 10.3389/fpsyt.2021.691251] [Medline: 34552514]
13. Verweel L, Newman A, Michaelchuk W, Packham T, Goldstein R, Brooks D. The effect of digital interventions on related health literacy and skills for individuals living with chronic diseases: a systematic review and meta-analysis. Int J Med Inform 2023 Sep;177:105114. [doi: 10.1016/j.ijmedinf.2023.105114] [Medline: 37329765]
14. Thomas TH, Go K, Go K, McKinley NJ, Dougherty KR, You K, et al. Empowerment through technology: a systematic evaluation of the content and quality of mobile applications to empower individuals with cancer. Int J Med Inform 2022 Jul;163:104782 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104782] [Medline: 35525126]
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. Dietrich D, Vorgeat H, Mazouri S, Ligier Q, Geissbuhler A. Engager les patients dans leurs soins à travers une application mobile: le projet Concerto. Rev Med Suisse 2018;617:1543-1547. [doi: 10.53738/revmed.2018.14.617.1543]
17. Federal Act on Data Protection. Fedlex. URL: https://www.fedlex.admin.ch/eli/cc/2022/491/en [accessed 2024-03-18]
18. CCER - obtain authorization for medical research on humans. Geneva Canton Ethics Board. URL: https://www.ge.ch/ccer-obtenir-autorisation-recherche-medicale-etre-humain [accessed 2024-03-18]
19. Pantaleoni J, Stevens L, Mailes E, Goad B, Longhurst C. Successful physician training program for large scale EMR implementation. Appl Clin Inform 2015 Dec 19;6(1):80-95 [FREE Full text] [doi: 10.4338/ACI-2014-09-CR-0076] [Medline: 25848415]

Abbreviations

BYOD: bring your own device
EMR: electronic medical record
HbA_{1c}: glycated hemoglobin
HIS: health information system
HUG: Hôpitaux Universitaires de Genève (University of Geneva Hospitals)
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
LDL-C: low-density lipoprotein cholesterol
PEP: patient empowerment program

Edited by A Benis; submitted 05.04.23; peer-reviewed by S Delaigue, KM Kuo; comments to author 10.06.23; revised version received 31.07.23; accepted 06.09.23; published 28.03.24.

Please cite as:
Dietrich D, Bornet dit Vorgeat H, Perrin Franck C, Ligier Q
A Mobile App (Concerto) to Empower Hospitalized Patients in a Swiss University Hospital: Development, Design, and Implementation Report
JMIR Med Inform 2024;12:e47914
URL: https://medinform.jmir.org/2024/1/e47914
doi:10.2196/47914
PMID:38546728

Scalable Approach to Consumer Wearable Postmarket Surveillance: Development and Validation Study

Richard M Yoo[1,*], MBI, PhD; Ben T Viggiano[1,*], BS; Krishna N Pundi[2,*], MD; Jason A Fries[1], PhD; Aydin Zahedivash[3], MBA, MD; Tanya Podchiyska[4], MS; Natasha Din[4], MBBS, MAS; Nigam H Shah[1,5,6], MBBS, PhD

1

2

3

4

5

6

[*]these authors contributed equally

Corresponding Author:
Richard M Yoo, MBI, PhD

Abstract

Background: With the capability to render prediagnoses, consumer wearables have the potential to affect subsequent diagnoses and the level of care in the health care delivery setting. Despite this, postmarket surveillance of consumer wearables has been hindered by the lack of codified terms in electronic health records (EHRs) to capture wearable use.

Objective: We sought to develop a weak supervision–based approach to demonstrate the feasibility and efficacy of EHR-based postmarket surveillance on consumer wearables that render atrial fibrillation (AF) prediagnoses.

Methods: We applied data programming, where labeling heuristics are expressed as code-based labeling functions, to detect incidents of AF prediagnoses. A labeler model was then derived from the predictions of the labeling functions using the Snorkel framework. The labeler model was applied to clinical notes to probabilistically label them, and the labeled notes were then used as a training set to fine-tune a classifier called Clinical-Longformer. The resulting classifier identified patients with an AF prediagnosis. A retrospective cohort study was conducted, where the baseline characteristics and subsequent care patterns of patients identified by the classifier were compared against those who did not receive a prediagnosis.

Results: The labeler model derived from the labeling functions showed high accuracy (0.92; F_1-score=0.77) on the training set. The classifier trained on the probabilistically labeled notes accurately identified patients with an AF prediagnosis (0.95; F_1-score=0.83). The cohort study conducted using the constructed system carried enough statistical power to verify the key findings of the Apple Heart Study, which enrolled a much larger number of participants, where patients who received a prediagnosis tended to be older, male, and White with higher CHA_2DS_2-VASc (congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category) scores (P<.001). We also made a novel discovery that patients with a prediagnosis were more likely to use anticoagulants (525/1037, 50.63% vs 5936/16,560, 35.85%) and have an eventual AF diagnosis (305/1037, 29.41% vs 262/16,560, 1.58%). At the index diagnosis, the existence of a prediagnosis did not distinguish patients based on clinical characteristics, but did correlate with anticoagulant prescription (P=.004 for apixaban and P=.01 for rivaroxaban).

Conclusions: Our work establishes the feasibility and efficacy of an EHR-based surveillance system for consumer wearables that render AF prediagnoses. Further work is necessary to generalize these findings for patient populations at other sites.

(JMIR Med Inform 2024;12:e51171) doi:10.2196/51171

KEYWORDS

consumer wearable devices; atrial fibrillation; postmarket surveillance; surveillance; monitoring; artificial intelligence; machine learning; natural language processing; NLP; wearable; wearables; labeler; heart; cardiology; arrhythmia; diagnose; diagnosis; labeling; classifier; EHR; electronic health record; electronic health records; consumer; consumers; device; devices; evaluation

Introduction

Background

Consumer-facing devices such as the Apple Watch [1] and Fitbit [2] now have the capability to notify users with a *prediagnosis* such as atrial fibrillation (AF). As these notifications may incentivize patients to seek follow-up medical care, wearables now have the potential to affect diagnosis rates and initiate cascades of medical care [3,4]. Although these devices undergo premarket validation to obtain Food and Drug Administration (FDA) clearance [5], limited information exists on their postmarket use and clinical utility.

To conduct *postmarket surveillance* on consumer wearables, electronic health records (EHRs) should capture wearable use, in particular those incidents where patients received prediagnosis notifications. However, EHRs are often built around medical diagnosis codes used for billing purposes [6,7], which do not contain terms for describing wearable use. Prescription wearables should have ordering information, but this does not capture how the wearables are used. Therefore, unstructured data such as clinical notes must be parsed to obtain the wearable use information.

Deep learning–based natural language processing (NLP) methods [8-10] have been shown to outperform traditional approaches on clinical note classification tasks [11,12]. However, these deep learning–based classifiers require large, hand-labeled training sets that are costly to generate. For EHR-based postmarket surveillance to be widely implemented, a scalable approach is necessary to reduce the labeling burden.

Objectives

We aimed to demonstrate the feasibility and efficacy of postmarket surveillance on consumer wearables that render AF prediagnoses. The first aim of this study was to evaluate the efficacy of a weakly supervised approach to heuristically generate labels for a training set. A *labeler model* derived from programmatically expressed heuristics probabilistically assigns labels to clinical notes regarding whether the note contains a mention of the patient receiving a prediagnosis from a wearable. The second aim was to evaluate the performance of a *classifier* fine-tuned on the training set labeled by the labeler model, which identifies mentions of an AF prediagnosis in a note. The third aim was to summarize the clinical characteristics of patients identified by the classifier and compare them to patients who were not alerted to a prediagnosis.

Methods

Cohort Identification

We used the Stanford Medicine Research Data Repository (STARR) data set [13], which contains EHR-derived records from the inpatient, outpatient, and emergency department visits at Stanford Health Care and the Lucile Packard Children's Hospital. We retrieved all clinical notes from the STARR data set that contain a mention of a wearable device (Textbox 1), resulting in 86,260 notes from 34,329 unique patients. Following the FDA guidance for pertinent cardiovascular algorithms [5], we excluded patients younger than 22 years of age when the note was written, leaving 78,323 notes from 30,133 unique patients. We further limited the data set to notes written on or after January 1, 2019, since the first consumer-facing AF detection feature became available in December 2018 [14]. The resulting cohort comprised 56,924 notes from 21,332 unique patients.

Textbox 1. Search terms for wearable devices.

Apple watch, iwatch, applewatch, fitbit, fit bit, fit-bit, galaxy watch, samsung watch, google watch, kardia, alivecor, alive cor, wearable, smart watch, and smartwatch

To evaluate the performance of the labeler model and the classifier, we constructed a test set by manually labeling 600 notes. Specifically, we randomly selected 600 unique patients and then selected 1 note for each patient that contained *action terms* (Textbox 2) in the vicinity (30 characters) of a wearable mention. This was to filter out nonrelevant wearable descriptions (eg, boilerplate texts recommending the use of wearables during meditation), so that resulting notes are enriched with relevant use cases.

Textbox 2. Action terms used to enrich sample relevance.

Alert, notify, warn, observe, identify, detect, note, record, capture, show, report, give, alarm, register, read, tell, have, had, see, saw, receive, get, got, notice, check, and confirm

These notes were then labeled independently by 2 data scientists, and differences were adjudicated by 2 physicians. A clinical note was labeled as positive when the patient received an automated AF notification from the wearable, or when the patient initiated an on-demand measurement (eg, electrocardiogram strip) that resulted in an AF prediagnosis. There were no instances where the 2 physicians disagreed on the label. The resulting test set contained 105 positive notes (prevalence=0.18).

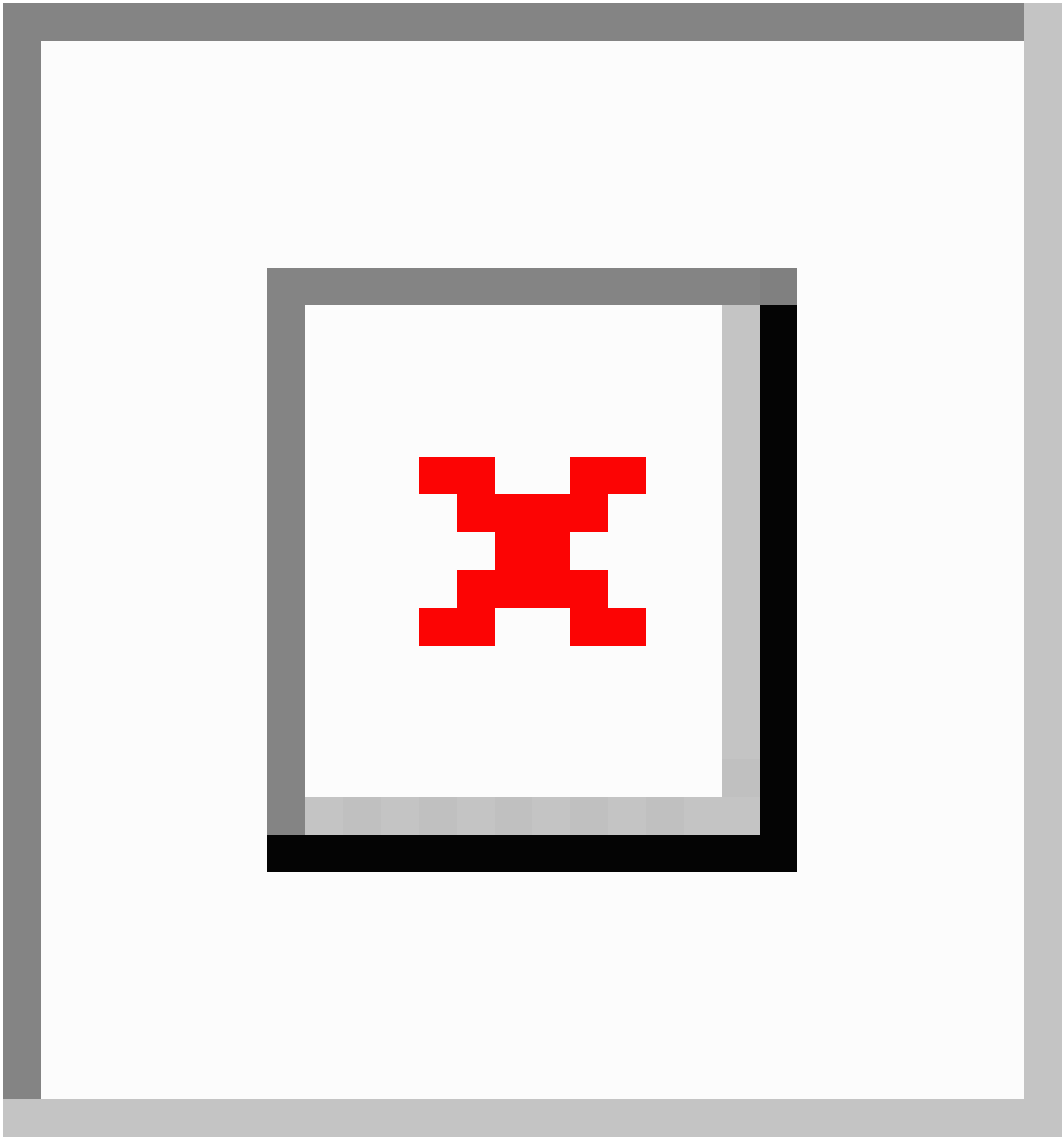
In addition to the test set, we prepared a development set of 600 notes that was used to aid the development of the labeler model. This set was manually labeled by a single data scientist, using a labeling guideline (Multimedia Appendix 1) that was developed as part of the test set generation. The development set contained 100 positive notes (prevalence=0.17).

Labeler Model Derivation

We then derived a labeler model that used weak supervision to probabilistically assign labels for the training set. Specifically,

as shown in Figure 1, we used data programming [15], where labeling heuristics are expressed as code-based *labeling functions*. Using the encoded heuristics, the labeling functions make predictions as to which label a clinical note should be assigned. Predictions from these labeling functions are then combined to develop a generative *labeler model*.

Figure 1. Labeler model generation process. Labeling heuristics were expressed as code-based labeling functions. Snorkel [16] then applied the labeling functions to the sample clinical notes and fit a generative model on the predictions of the labeling functions. The resulting labeler model probabilistically assigns a label to a clinical note based on whether the note mentions the patient receiving an AF prediagnosis from the wearable device. AF: atrial fibrillation.



We used the Snorkel framework [16] to implement data programming. A preprocessing framework [17] was applied to our notes to split them into sentences using the spaCy [18] framework, with a specialized tokenizer to recognize terms specific to medical literature. Thus parsed grammatical information was made available to the labeling functions as metadata.

We then used the development set to understand how the AF prediagnosis was described, and we expressed each pattern as a labeling function. The development process was iterative, where the Snorkel framework allowed us to observe the predictive values of the labeling functions on development set records. Each function could then be further optimized to reduce the differences between predictive values and actual labels, leading to overall performance improvement on the development

set. Textbox 3 shows all the terms that were identified as denoting AF. Negations were properly handled.

Once developed, we applied the labeling functions on the samples and then instructed Snorkel to fit a generative model on the output. Specifically, we used 10-fold cross-validation on the test set and chose the labeler model with the best F_1-score. This model was then applied to the entire corpus of 56,924 notes to probabilistically assign labels.

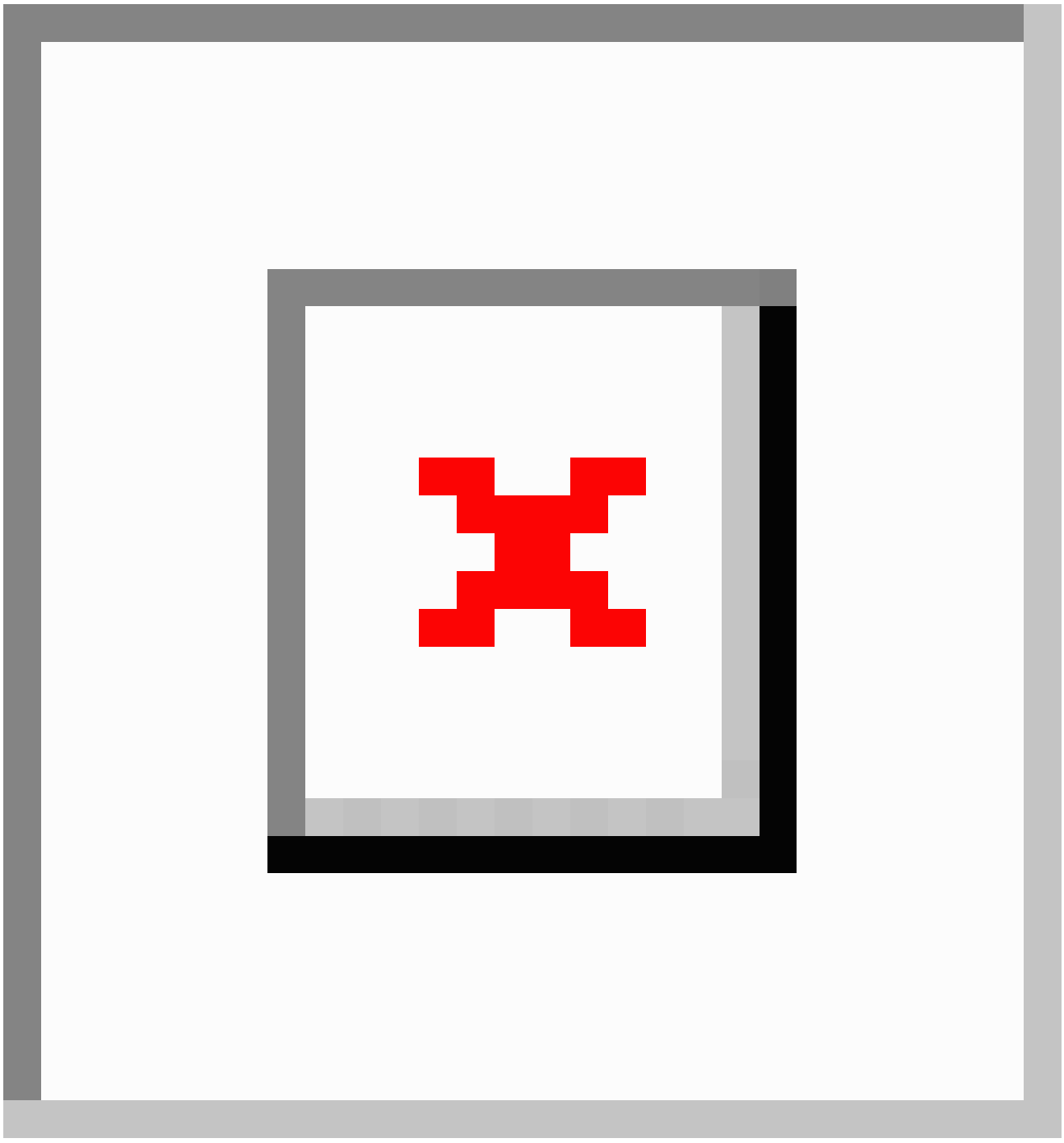
Textbox 3. Terms denoting atrial fibrillation.

Af, afib, a-fib, a.fib, arrhythmia, paf, atrial fibrillation, a. fib, a fib, atrial fib, atrial arrhythmia, irregular heartbeat, irregular hr, irregular rhythm, irregular pulse, irreg hr, irregular heart beat, irregular heart rhythm, irregular heart rate, irreg heart rhythm, irreg heart beat, irreg heart rate, abnormal ekg rhythm, paroxysmal atrial fibrillation, a. fib, a - fib, pafib, abnormal heart rhythm, abnormal rhythm, abnormal HR, and arrhythmia

Classifier Fine-Tuning

Notes that were probabilistically labeled by the labeler model were then used to fine-tune a large, NLP-based classifier: Clinical-Longformer [12] (Figure 2). The resulting classifier takes plain note text as the input and classifies the note as positive (ie, includes mention of a patient receiving an AF notification, or patient-initiated cardiac testing or electrocardiogram resulting in an AF prediagnosis) or negative. When a classifier is tuned on the labeler model output, it enables generalizing beyond the labeling heuristics encoded in the labeling functions, such that the classifier can recognize more patterns.

Figure 2. Classifier generation process. The labeler model was used to probabilistically assign labels for a large number of unlabeled clinical notes, which were then used to fine-tune a classifier to detect whether a patient received an AF prediagnosis from a wearable device. AF: atrial fibrillation; NLP: natural language processing.



Specifically, we fine-tuned the pretrained Clinical-Longformer for the sequence classification task, with varying training set sizes. For a single fine-tuning run, we chose the snapshot with the best F_1-score on the test set as the representative. The Adam optimizer was used, with the learning rate ramping up to 1×10^{-5} followed by linear decay over 3 epochs. Clinical-Longformer has a maximum input length of 4096 subword tokens: 94% (53,509/56,924) of our notes fit this criterion, and notes with more tokens were trimmed. Fine-tuning other NLP-based classifiers (eg, ClinicalBERT [11], which takes a smaller number of input tokens [512 or fewer]) resulted in abysmal performance numbers (F_1-score=0.21), hinting that they could not be properly fine-tuned on our lengthy clinical notes.

The test set was never presented to the classifier during the fine-tuning process. Since our data set was highly skewed toward negative samples, we stratified the training set to maintain a 1:2 ratio between the positive and negative notes. All samples were chosen randomly.

The classifier with the best F_1-score was then run across the entire set of 56,924 clinical notes to identify all incidents of AF prediagnoses.

Retrospective Cohort Study

Using the classifier, we identified patients who received an AF prediagnosis and performed 3 retrospective cohort studies comparing the characteristics of patients who received a prediagnosis to those who did not, using the same STARR data set.

First, we considered all the patients in the cohort regardless of their prior AF diagnosis. We compared the demographics, CHA_2DS_2-VASc (congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category) [19] score, and its related comorbidities on the date the index note was created. We defined the oldest note with a prediagnosis as the index note since it was the most likely to drive downstream medical intervention. When a patient had not received any prediagnosis, the oldest note with mention of a wearable was chosen as the index.

Second, we focused on patients who did not have a prior AF diagnosis. A patient was filtered out if the patient had received an AF diagnosis, defined as an ambulatory or inpatient encounter with SNOMED code 313217 and its descendants, prior to the index note. We then compared the same demographics and comorbidities between those who received a prediagnosis and those who did not, on the date the index note was created.

Lastly, we further confined the analysis to patients who received a clinician-assigned AF diagnosis within 60 days from the index note. Same as before, we excluded patients who had a prior AF diagnosis before the index note. Patients were then grouped based on whether they had received an AF prediagnosis from a wearable and characterized on the date they received the index AF diagnosis. In addition to the demographics and comorbidities, we also compared anticoagulant medication (Textbox 4), rhythm management medication (Textbox 5), and cardioversion rates between the 2 groups. Only the index prescription and procedure that took place within 60 days from the index diagnosis were considered.

Textbox 4. Anticoagulant medications analyzed in this study.

Warfarin

Direct oral anticoagulants

- Apixaban, dabigatran, rivaroxaban, edoxaban, and betrixaban

Textbox 5. Rhythm management medications analyzed in this study.

Class I antiarrhythmics

- Propafenone, disopyramide, quinidine, mexiletine, and flecainide

Class II antiarrhythmics

- Metoprolol, carvedilol, labetalol, nadolol, propranolol, carteolol, penbutolol, pindolol, atenolol, betaxolol, bisoprolol, esmolol, nebivolol, and timolol

Class III antiarrhythmics

- Sotalol and dofetilide

Class IV antiarrhythmics

- Verapamil, diltiazem, nicardipine, amlodipine, felodipine, nifedipine, isradipine, and nisoldipine

Others

- Digoxin

Statistical Analysis

When compiling patient race and ethnicity information, we used the 5 categories of race defined by the US Census and denoted Hispanic as a dedicated ethnicity. A total of 11.12% (2371/21,327) of the patients were missing race and ethnicity information, so we categorized them as belonging to the *undisclosed* category.

For hypothesis testing, we used the 1-tailed Welch *t* test for continuous variables and χ^2 test for categorical variables. One-tailed tests were chosen over 2-tailed tests since clinical contexts helped establish the comparison direction, providing for a stricter analysis. Statistical analysis was performed using *Pandas* [20] 1.3.0 and *SciPy* [21] 1.7.0, running on Python 3.9.6 configured through Conda 4.5.11.

Ethical Considerations

The STARR data set is derived from consented patients only. Patients were not compensated for participation. Data analyzed in this study were not deidentified, but its analysis was conducted in a HIPAA (Health Insurance Portability and Accountability Act)–compliant, high-security environment. The Stanford University Institutional Review Board approved this study (62865).

Results

Labeler Model Performance

In total, 8 labeling functions were developed. Most (7/8, 88%) labeling functions used the grammatical information present in the metadata, whereas 1 (12%) used a simple dictionary-based lookup. Table 1 provides the performance of each labeling function, followed by the combined labeler model.

Since each labeling function was geared toward identifying positive samples that follow a specific pattern, each labeling function exhibited substantially higher precision than recall. By combining these labeling functions into 1 generative labeler model, we improved recall (0.72). The high labeler model accuracy (0.92) also showed that the model correctly classified negative samples. After running the labeler model on the set of 56,924 clinical notes, 5829 notes were flagged as positive, a substantial increase from the 105 positive notes identified through manual labeling.

Table . Labeling function (LF) and labeler model performance[a].

Function or model	Target pattern	Example	Precision[b]	Recall[c]	F_1-score[d]	Accuracy[e]
LF1	Simple dictionary lookup	"AF" and "wearable" and "notification"	0.90	0.33	0.51	0.87
LF2	AF[f]+verb+prep[g]+wearable	"AF noted on wearable"	0.78	0.12	0.24	0.84
LF3	Wearable+verb+AF	"Wearable notified AF"	0.91	0.42	0.55	0.89
LF4	Verb+wearable+verb+AF	"Observed wearable showing AF"	0.85	0.14	0.29	0.85
LF5	Verb+AF+prep+wearable	"Received AF from wearable"	0.81	0.15	0.31	0.85
LF6	Verb+event+prep+wearable+AF	"Got notification from wearable of AF"	0.67	0.02	0.20	0.83
LF7	Event+prep+wearable+AF	"Notified on wearable of AF"	0.74	0.10	0.27	0.84
LF8	Wearable+subject+verb+AF	"Per wearable, patient had AF"	*0.96*	0.22	0.38	0.86
Labeler model	N/A[h]	N/A	0.84	*0.72*	*0.77*	*0.92*

[a]Averages taken from 10-fold cross-validation on the test set of 600 manually labeled notes. Italic numbers indicate the best observed performance for each metric.

[b]Precision = true positive / (true positive + false positive).

[c]Recall = true positive / (true positive + false negative).

[d]F_1-score = 2 × precision × recall / (precision + recall).

[e]Accuracy = (true positive + true negative) / (positive + negative).

[f]AF: atrial fibrillation.

[g]Prep: preposition.

[h]N/A: not available.

Classifier Performance

Here, we report the performance of the classifier that was fine-tuned using the clinical notes labeled by the labeler model. Table 2 shows the average performance of the classifier on the test set, across varying training set sizes. The training set size was capped at 15,000 to maintain the 1:2 positive-to-negative ratio (the labeler model labeled 5829 notes as positive). Regardless of the training set size, the test set was excluded from the input to the fine-tuning process.

Table . Classifier performance across varying training set sizes[a].

Training set size	Precision[b]	Recall[c]	F_1-score[d]	Accuracy[e]
600	0.37	0.68	0.48	0.73
5000	0.79	*0.85*	0.81	0.93
10,000	0.84	0.81	*0.83*	*0.94*
15,000	*0.85*	0.81	*0.83*	*0.94*

[a]For each training set, average values are reported across 3 runs with different random seeds. For each run, the classifier snapshot with highest F_1-score was used. Italic numbers indicate the best performance observed for each metric.

[b]Precision = true positive / (true positive + false positive).

[c]Recall = true positive / (true positive + false negative).

[d]F_1-score = 2 × precision × recall / (precision + recall).

[e]Accuracy = (true positive + true negative) / (positive + negative).

Table 2 demonstrates how classifier performance benefits from the weakly supervised approach. In particular, a training set size of 600 emulated the hypothetical scenario where the size of the training set is limited due to manual labeling overhead. Such a small data set was not enough to adequately fine-tune Clinical-Longformer (F_1-score=0.48).

As the training set size increased, the classifier obtained better performance, reaching the best average F_1-score of 0.83. When compared to the labeler model in Table 1 (recall=0.72), the classifier significantly improved recall (0.81), demonstrating that the classifier managed to generalize beyond the rules specified by the labeling functions.

Figures 3 and 4 show the comparisons of the best-performing (by F_1-score) classifiers from each training set size.

Figure 3. Classifier receiver operating characteristic (ROC) curve across varying training set sizes. For each training set, the best-performing (by F_1-score) run was chosen among 3 runs with different random seeds. For each run, the best-performing classifier snapshot was chosen.

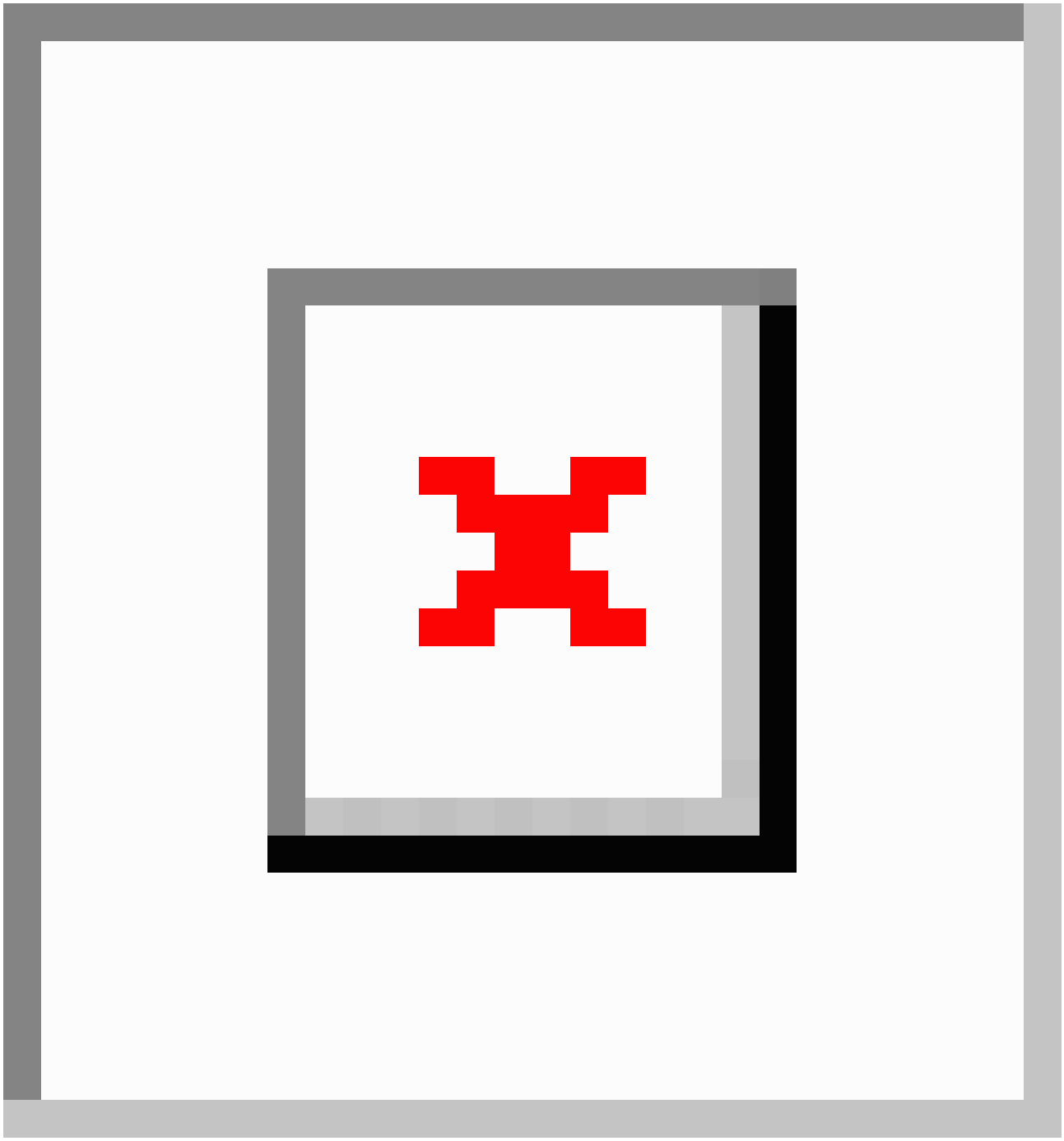
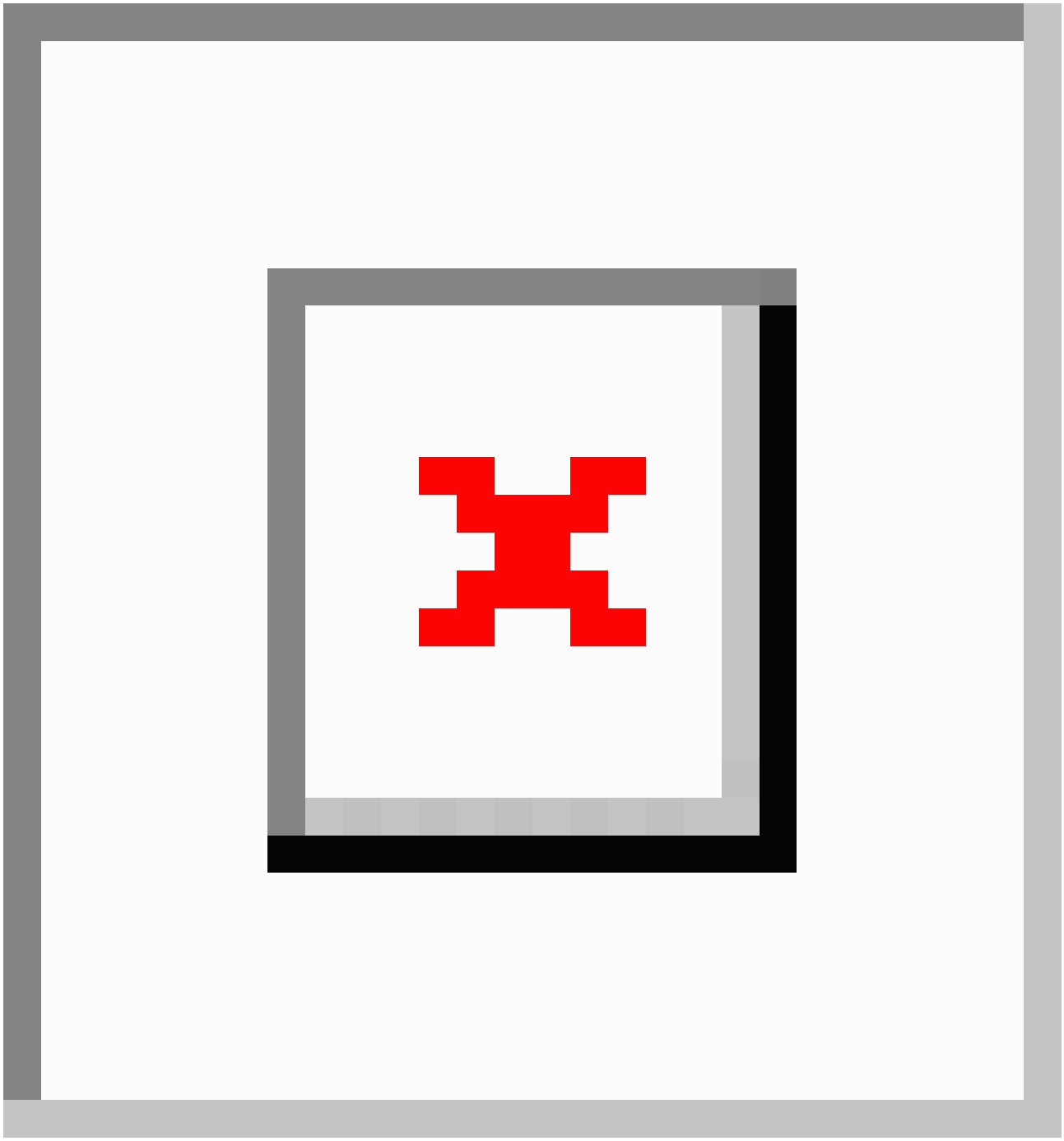


Figure 4. Classifier precision-recall curve across varying training set sizes. For each training set, the best-performing (by F_1-score) run was chosen among 3 runs with different random seeds. For each run, the best-performing classifier snapshot was chosen.



The receiver operating characteristic curve (Figure 3) shows that even the best classifier with a training set size of 600 performed worse than classifiers from larger data set sizes. In the precision-recall curve (Figure 4), the classifier lost substantial precision for small gains in recall, further hinting that the classifier was not properly trained.

Across all training set sizes and runs, the best-performing classifier achieved an F_1-score of 0.85 (accuracy=0.95). Running this classifier on 56,924 clinical notes identified 6515 notes as containing an AF prediagnosis across 2279 unique patients.

Cohort Study: All Patients

Table 3 summarizes the characteristics of the entire cohort regardless of their prior AF diagnosis, reflecting the characteristics of generic patients that used wearables. In all, 5 patients were missing sex information and were not included in the analysis.

Table . Characteristics of all patients[a].

Characteristics			With a prediagnosis (n=2279)	Without a prediagnosis (n=19,048)	*P* value
Demographics					
	Age (y), mean (SD)		63.85 (14.21)	53.53 (16.70)	<.001[b]
	Race and ethnicity, n (%)				<.001[b]
		Asian	295 (12.94)	3143 (16.5)	
		Black	53 (2.33)	619 (3.25)	
		Hispanic	96 (4.21)	1731 (9.09)	
		White	1613 (70.78)	11,240 (59.01)	
		Others	13 (0.57)	153 (0.8)	
		Undisclosed	209 (9.17)	2162 (11.35)	
	Sex, n (%)				<.001[b]
		Male	1384 (60.73)	7739 (40.63)	
		Female	895 (39.27)	11,309 (59.37)	
Comorbidities, n (%)					
	Congestive heart failure		341 (14.96)	1434 (7.53)	<.001[b]
	Hypertension		1267 (55.59)	6796 (35.68)	<.001[b]
	Diabetes mellitus		101 (4.43)	1018 (5.34)	.07
	Vascular disease		251 (11.01)	1582 (8.31)	<.001[b]
CHA_2DS_2-VASc[c] score, mean (SD)			2.12 (1.55)	1.61 (1.35)	<.001[b]

[a]Measured on the date of the index note.

[b]Statistically significant at α=.05.

[c]CHA_2DS_2-VASc: congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category.

Patients who received an AF prediagnosis from a wearable tended to be older, with more comorbidities except for diabetes mellitus. White and male individuals constituted a larger portion of patients with a prediagnosis, who also exhibited higher CHA_2DS_2-VASc scores.

Cohort Study: Patients Without a Prior AF Diagnosis

Table 4 then compares the characteristics of patients who had no AF diagnosis prior to the index note, highlighting the efficacy of wearables on the undiagnosed population.

Table . Characteristics of patients without a prior atrial fibrillation diagnosis[a].

Characteristics			With a prediagnosis (n=1037)	Without a prediagnosis (n=16,560)	*P* value
Demographics					
	Age (y), mean (SD)		60.16 (15.65)	51.54 (16.28)	<.001[b]
	Race and ethnicity, n (%)				<.001[b]
		Asian	127 (12.25)	2890 (17.45)	
		Black	28 (2.7)	553 (3.34)	
		Hispanic	55 (5.3)	1598 (9.65)	
		White	723 (69.72)	9414 (56.85)	
		Others	3 (0.29)	136 (0.82)	
		Undisclosed	101 (9.74)	1969 (11.89)	
	Sex, n (%)				<.001[b]
		Male	595 (57.38)	6241 (37.69)	
		Female	442 (42.62)	10,319 (62.31)	
Comorbidities, n (%)					
	Congestive heart failure		85 (8.2)	696 (4.2)	<.001[b]
	Hypertension		461 (44.46)	5082 (30.69)	<.001[b]
	Diabetes mellitus		42 (4.05)	805 (4.86)	.27
	Vascular disease		95 (9.16)	1090 (6.58)	.002[b]
CHA_2DS_2-VASc[c] score, mean (SD)			1.78 (1.44)	1.46 (1.23)	<.001[b]

[a]Measured on the date of the index note.

[b]Statistically significant at α=.05.

[c]CHA_2DS_2-VASc: congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category.

These patients exhibited similar characteristics to the overall cohort, where those who received an AF prediagnosis tended to be older, White, and male, with more comorbidities except for diabetes mellitus. In particular, 50.63% (525/1037) of the patients who received a prediagnosis had CHA_2DS_2-VASc scores of 2 or higher, warranting anticoagulation therapy [22]. In contrast, among the patients without a prediagnosis, only 35.85% (5936/16,560) had CHA_2DS_2-VASc scores of 2 or higher.

Cohort Study: Patients With a Clinician-Assigned AF Diagnosis

Among those patients who did not have a prior AF diagnosis, 29.41% (305/1037) of the patients with a wearable-assigned prediagnosis received a clinician-assigned AF diagnosis within 60 days from the index prediagnosis. The average duration from prediagnosis to diagnosis was 4.74 days. In contrast, only 1.58% (262/16,560) of those patients without a prediagnosis received a clinician-assigned AF diagnosis.

Table 5 compares the clinical characteristics of those patients who received an AF diagnosis, based on whether they had received a wearable-assigned prediagnosis prior to the diagnosis.

None of the patient characteristics reported in Table 5 differed significantly between those with an AF prediagnosis and those without (all *P*>.05). However, anticoagulant prescriptions differed based on AF prediagnoses, where more patients with a prediagnosis were prescribed apixaban and rivaroxaban.

Table . Characteristics of patients with a clinician-assigned atrial fibrillation diagnosis[a].

Characteristics		With a prediagnosis (n=305)	Without a prediagnosis (n=262)	*P* value
Demographics				
Age (y), mean (SD)		64.45 (14.16)	63.65 (14.29)	.75
Race and ethnicity, n (%)				.21
	Asian	35 (11.48)	27 (10.31)	
	Black	6 (1.97)	11 (4.20)	
	Hispanic	10 (3.28)	15 (5.73)	
	White	218 (71.48)	175 (66.79)	
	Others	1 (0.33)	5 (1.91)	
	Undisclosed	35 (11.48)	29 (11.07)	
Sex, n (%)				.86
	Male	193 (63.28)	163 (62.21)	
	Female	112 (36.72)	99 (37.79)	
Comorbidities, n (%)				
Congestive heart failure		14 (4.59)	21 (8.02)	.13
Hypertension		111 (36.39)	109 (41.6)	.24
Diabetes mellitus		10 (3.28)	4 (1.53)	.29
Vascular disease		24 (7.87)	30 (11.45)	.19
CHA_2DS_2-VASc[b] score, mean (SD)		1.76 (1.49)	1.81 (1.39)	.36
Diagnosis subtype, n (%)				.40
Generic		230 (75.41)	213 (81.3)	
Chronic		2 (0.66)	1 (0.38)	
Paroxysmal		68 (22.3)	45 (17.18)	
Persistent		5 (1.64)	3 (1.15)	
Anticoagulant, n (%)				
Warfarin		1 (0.33)	3 (1.15)	.51
Direct oral anticoagulants				
	Apixaban	76 (24.92)	39 (14.89)	.004[c]
	Rivaroxaban	29 (9.51)	10 (3.82)	.01[c]
Rhythm management, n (%)				
Class I antiarrhythmics				
	Propafenone	7 (2.3)	2 (0.76)	.26
	Flecainide	17 (5.57)	8 (3.05)	.21
Class II antiarrhythmics				
	Metoprolol	50 (16.39)	45 (17.18)	.89
	Carvedilol	1 (0.33)	3 (1.15)	.51
	Labetalol	6 (1.97)	4 (1.53)	.94
	Atenolol	3 (0.98)	5 (1.91)	.57
Class IV antiarrhythmics				
	Verapamil	3 (0.98)	2 (0.76)	>.99
	Diltiazem	15 (4.92)	9 (3.44)	.51

Characteristics		With a prediagnosis (n=305)	Without a prediagnosis (n=262)	*P* value
	Amlodipine	4 (1.31)	3 (1.15)	>.99
Others				
	Digoxin	3 (0.98)	1 (0.38)	.73
Procedures, n (%)				
Cardioversion		30 (9.84)	14 (5.34)	.07

[a]Measured on the date of the index atrial fibrillation diagnosis. Medications that were not prescribed are omitted.

[b]CHA_2DS_2-VASc: congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category.

[c]Statistically significant at α=.05.

Discussion

Principal Findings

In this study, we applied a weak supervision–based approach to demonstrate the feasibility and efficacy of an EHR-based postmarket surveillance system for consumer wearables that render AF prediagnoses.

We first derived a labeler model from labeling heuristics expressed as labeling functions, which showed high accuracy (0.92; F_1-score=0.77) on the test set. We then fine-tuned a classifier on labeler model output, to accurately identify AF prediagnoses (0.95; F_1-score=0.83).

Further, using the classifier output, we identified patients who received an AF prediagnosis from a wearable and conducted a retrospective analysis to compare the baseline characteristics and subsequent clinical treatment of these patients against those who did not receive a prediagnosis.

Across the entire cohort, patients with a prediagnosis were older with more comorbidities. The race and sex composition of these patients also differed from those who did not receive a prediagnosis (*P*<.001).

Focusing on the subgroup of patients without a prior AF diagnosis (Table 4), we observed that a higher percentage of patients (525/1037, 50.63% vs 5936/16,560, 35.85%) who received a wearable-assigned prediagnosis exhibited CHA_2DS_2-VASc scores that warranted a recommendation for anticoagulation therapy [22]. This increased likelihood for anticoagulation therapy could be attributed to an early prediagnosis from the wearable.

In the same subgroup, patients who received a prediagnosis were 18.61 times more likely to receive a clinician-assigned AF diagnosis than those who did not. The existence of a prediagnosis was not correlated with patient demographics, comorbidities, or AF subtype at the index diagnosis (Table 5) but did correlate with anticoagulant prescription, where patients with an AF prediagnosis were more frequently prescribed apixaban (*P*=.004) and rivaroxaban (*P*=.01).

Comparison With Prior Work

Given that more consumer wearables will be introduced with increasing prediagnostic capabilities, a surveillance framework for wearable devices is urgently needed to properly assess their impact on downstream health care [3,4]. However, publications sponsored by wearable vendors focused mostly on ascertaining the accuracy of the prediagnostic algorithm itself [1,2].

On the other hand, publications that sought to conduct postmarket surveillance relied solely on manual chart review [3,4], which is hard to scale. In a prior study on wearable notifications, clinician review of 534 clinical notes yielded only 41 patients with an AF prediagnosis [3]. With a weakly supervised approach, our clinician review of 600 notes (ie, the test set) allowed the subsequent identification of 2279 patients with a prediagnosis.

Such an improvement in recall enhanced the statistical power of our analysis. First, our cohort study findings that showed patients with an AF prediagnosis tended to be older, male, and White with higher CHA_2DS_2-VASc scores matches the key findings of the Apple Heart Study [1], which enrolled a much larger number of participants (n=419,297). Second, we were able to make a novel discovery in that a wearable-assigned prediagnosis increases the likelihood of patients receiving anticoagulation therapy and an eventual AF diagnosis, and we identified statistically meaningful anticoagulant prescription differences.

Prior work has applied various methods of weakly supervised learning to some form of medical surveillance [16,17,23-25]. Most relevantly, Callahan et al [23] implemented a surveillance framework for hip implants, and Sanyal et al [25] implemented one for insulin pumps. To the best of our knowledge, however, our work is the first to apply a weakly supervised approach to consumer wearable surveillance. Without prescription records, consumer wearable surveillance can be challenging to scale.

Limitations

We acknowledge that the STARR data set is confined to a small health care system in a single geographic region, which is known [13] to serve populations with higher percentages of male, White, and older individuals. We recommend other institutions to monitor their patient population by developing their own surveillance framework using our weakly supervised methodology. In fact, work is already underway to adapt this approach for use at Palo Alto Veterans Affairs.

We could not establish causality between prediagnoses and patient characteristics. The fact that patients who are older, with more comorbidities; White; and male had a higher likelihood

of receiving an AF prediagnosis may very well reflect that they are health conscious and use wearables more frequently.

Conclusions

By providing prediagnoses, consumer wearables have the potential to affect subsequent diagnoses and downstream health care. Postmarket surveillance of wearables is necessary to understand the impact but is hindered by the lack of codified terms in EHRs to capture wearable use. By applying a weakly supervised methodology to efficiently identify wearable-assigned AF prediagnoses from clinical notes, we demonstrate that such a surveillance system could be built.

The cohort study conducted using the constructed system carried enough statistical power to verify the key findings of the Apple Heart Study, which enrolled a much larger number of patients, where patients who received a prediagnosis tended to be older, male, and White with higher CHA_2DS_2-VASc scores. We also made a novel discovery in that a prediagnosis from a wearable increases the likelihood for anticoagulant prescription and an eventual AF diagnosis. At the index diagnosis, the existence of a prediagnosis from a wearable did not distinguish patients based on clinical characteristics but did correlate with anticoagulant prescription.

Our work establishes the feasibility and efficacy of an EHR-based surveillance system for consumer wearable devices. Further work is necessary to generalize these findings for patient populations at other sites.

Authors' Contributions

RMY, BTV, KNP, JAF, and NHS contributed to concept and design. JAF contributed to the acquisition of data. RMY, BTV, KNP, JAF, AZ, TP, and ND contributed to the analysis and interpretation of data. RMY and BTV contributed to the drafting of the manuscript. RMY, KNP, JAF, AZ, TP, ND, and NHS contributed to critical revision of the manuscript for important intellectual content. RMY contributed to statistical analysis. NHS contributed to the provision of patients or study materials, obtaining funding, and supervision. JAF and NHS contributed to administrative, technical, or logistic support.

Conflicts of Interest

KNP receives research grants from the American Heart Association and the American College of Cardiology and is a consultant for Evidently and 100Plus. JAF is a research consultant for Snorkel AI.

Multimedia Appendix 1
Labeling guideline developed as part of the test set generation.
[DOCX File, 46 KB - medinform_v12i1e51171_app1.docx]

References

1. Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 2019 Nov 14;381(20):1909-1917. [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
2. Lubitz SA, Faranesh AZ, Selvaggi C, et al. Detection of atrial fibrillation in a large population using wearable devices: the Fitbit Heart Study. Circulation 2022 Nov 8;146(19):1415-1424. [doi: 10.1161/CIRCULATIONAHA.122.060291] [Medline: 36148649]
3. Wyatt KD, Poole LR, Mullan AF, Kopecky SL, Heaton HA. Clinical evaluation and diagnostic yield following evaluation of abnormal pulse detected using Apple Watch. J Am Med Inform Assoc 2020 Jul 1;27(9):1359-1363. [doi: 10.1093/jamia/ocaa137] [Medline: 32979046]
4. Feldman K, Duncan RG, Nguyen A, et al. Will Apple devices' passive atrial fibrillation detection prevent strokes? estimating the proportion of high-risk actionable patients with real-world user data. J Am Med Inform Assoc 2022 May 11;29(6):1040-1049. [doi: 10.1093/jamia/ocac009] [Medline: 35190832]
5. Device classification under section 513(F)(2)(De Novo). US Food and Drug Administration. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN180044 [accessed 2023-03-05]
6. ICD-10. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/Medicare/Coding/ICD10 [accessed 2023-04-28]
7. List of CPT/HCPCS codes. Centers for Medicare & Medicaid Services. URL: https://www.cms.gov/medicare/fraud-and-abuse/physicianselfreferral/list_of_codes [accessed 2023-04-28]
8. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, et al, editors. Advances in Neural Information Processing Systems 30 (NIPS 2017): Curran Associates Inc; 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [accessed 2023-03-05]
9. 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, Volume 1 (Long and Short Papers): Association for Computational Linguistics; 2019:4171-4186. [doi: 10.18653/v1/N19-1423]
10. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv. Preprint posted online on Apr 10, 2020. [doi: 10.48550/arXiv.2004.05150]

11. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on Apr 10, 2019. [doi: 10.48550/arXiv.1904.05342]
12. 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]
13. Datta S, Posada J, Olson G, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv. Preprint posted online on Mar 17, 2020. [doi: 10.48550/arXiv.2003.10534]
14. ECG app and irregular heart rhythm notification available today on Apple Watch. Apple. 2018 Dec 6. URL: https://www.apple.com/newsroom/2018/12/ecg-app-and-irregular-heart-rhythm-notification-available-today-on-apple-watch/ [accessed 2023-3-5]
15. Ratner A, De Sa C, Wu S, Selsam D, Ré C. Data programming: creating large training sets, quickly. In: NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems: Curran Associates Inc; 2016:3574-3582. [doi: 10.5555/3157382.3157497]
16. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. VLDB J 2020;29(2):709-730. [doi: 10.1007/s00778-019-00552-1] [Medline: 32214778]
17. Fries JA, Steinberg E, Khattar S, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. Nat Commun 2021 Apr 1;12(1):2017. [doi: 10.1038/s41467-021-22328-4] [Medline: 33795682]
18. Industrial-strength natural language processing in Python. spaCy. URL: https://spacy.io/ [accessed 2023-03-05]
19. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro Heart Survey on Atrial Fibrillation. Chest 2010 Feb;137(2):263-272. [doi: 10.1378/chest.09-1584] [Medline: 19762550]
20. McKinney W. Data structures for statistical computing in Python. Presented at: 9th Python in Science Conference (SciPy 2010); Jun 28 to Jul 3, 2010; Austin, Texas p. 56-61. [doi: 10.25080/Majora-92bf1922-00a]
21. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):261-272. [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
22. The revised ACC/AHA/HRS guidelines for the management of patients with atrial fibrillation. American College of Cardiology. 2014 Oct 29. URL: https://www.acc.org/latest-in-cardiology/articles/2014/10/14/11/02/the-revised-acc-aha-hrs-guidelines-for-the-management-of-patients-with-atrial-fibrillation [accessed 2023-03-12]
23. Callahan A, Fries JA, Ré C, et al. Medical device surveillance with electronic health records. NPJ Digit Med 2019 Sep 25;2:94. [doi: 10.1038/s41746-019-0168-z] [Medline: 31583282]
24. Datta S, Roberts K. Weakly supervised spatial relation extraction from radiology reports. JAMIA Open 2023 Apr 22;6(2):ooad027. [doi: 10.1093/jamiaopen/ooad027] [Medline: 37096148]
25. Sanyal J, Rubin D, Banerjee I. A weakly supervised model for the automated detection of adverse events using clinical notes. J Biomed Inform 2022 Feb;126:103969. [doi: 10.1016/j.jbi.2021.103969] [Medline: 34864210]

Abbreviations

AF: atrial fibrillation
CHA_2DS_2-VASc: congestive heart failure, hypertension, age ≥75 years, diabetes, stroke, vascular disease, age 65-74 years, sex category
EHR: electronic health record
FDA: Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
NLP: natural language processing
STARR: Stanford Medicine Research Data Repository

Edited by C Lovis; submitted 29.07.23; peer-reviewed by D Teo, L Wu; revised version received 15.01.24; accepted 04.02.24; published 04.04.24.

Please cite as:
Yoo RM, Viggiano BT, Pundi KN, Fries JA, Zahedivash A, Podchiyska T, Din N, Shah NH
Scalable Approach to Consumer Wearable Postmarket Surveillance: Development and Validation Study
JMIR Med Inform 2024;12:e51171
URL: https://medinform.jmir.org/2024/1/e51171
doi:10.2196/51171

Addressing Hospital Overwhelm During the COVID-19 Pandemic by Using a Primary Health Care–Based Integrated Health System: Modeling Study

Jiaoling Huang[1], PhD; Ying Qian[2,3], PhD; Yuge Yan[1]; Hong Liang[4], PhD; Laijun Zhao[2,3], PhD

1

2

3

4

Corresponding Author:
Ying Qian, PhD

Abstract

Background: After strict COVID-19–related restrictions were lifted, health systems globally were overwhelmed. Much has been discussed about how health systems could better prepare for future pandemics; however, primary health care (PHC) has been largely ignored.

Objective: We aimed to investigate what combined policies PHC could apply to strengthen the health care system via a bottom-up approach, so as to better respond to a public health emergency.

Methods: We developed a system dynamics model to replicate Shanghai's response when COVID-19–related restrictions were lifted. We then simulated an alternative PHC-based integrated health system and tested the following three interventions: first contact in PHC with telemedicine services, recommendation to secondary care, and return to PHC for recovery.

Results: The simulation results showed that each selected intervention could alleviate hospital overwhelm. Increasing the rate of first contact in PHC with telemedicine increased hospital bed availability by 6% to 12% and reduced the cumulative number of deaths by 35%. More precise recommendations had a limited impact on hospital overwhelm (<1%), but the simulation results showed that underrecommendation (rate: 80%) would result in a 19% increase in cumulative deaths. Increasing the rate of return to PHC from 5% to 20% improved hospital bed availability by 6% to 16% and reduced the cumulative number of deaths by 46%. Moreover, combining all 3 interventions had a multiplier effect; bed availability increased by 683%, and the cumulative number of deaths dropped by 75%.

Conclusions: Rather than focusing on the allocation of medical resources in secondary care, we determined that an optimal PHC-based integrated strategy would be to have a 60% rate of first contact in PHC, a 110% recommendation rate, and a 20% rate of return to PHC. This could increase health system resilience during public health emergencies.

(JMIR Med Inform 2024;12:e54355) doi:10.2196/54355

KEYWORDS

hospital overwhelm; primary health care; modeling study; policy mix; pandemic; model; simulation; simulations; integrated; health system; hospital; hospitals; management; service; services; health systems; develop; development; bed; beds; overwhelm; death; deaths; mortality; primary care

Introduction

The World Health Organization (WHO) announced the end of the COVID-19 public health emergency of international concern on May 5, 2023. Over the past 3 years, the COVID-19 epidemic has resulted in more than 765 million infections and 6.92 million deaths globally and has involved ongoing outbreaks, infection control via restrictions, the lifting of restrictions, and large-scale infections [1]. The limitations of health care systems worldwide regarding the response to mass infections and admissions have been exposed, and these limitations exist in countries classified as high-performing and resilient countries, as well as in resource-limited countries [2-4]. Although most governments have prudently considered the appropriate time to relax restriction policies, health care systems have unavoidably been overwhelmed, and some even collapsed once restrictions were lifted [5].

Much has been discussed regarding how health care systems could have better prepared for COVID-19 and how to prepare for future pandemics. Topics of discussion include adequate health care workforces and facilities [6], better intensive care unit capacity [7], early intervention to avoid local transmission

[8], and the broader application of telemedicine [9]. Some scholars have advocated for the integration and coordination of the health system, including public health and clinical medicine. The role of primary health care (PHC) in COVID-19 management has received attention [10-13], but this attention is obviously insufficient when compared with the attention given to the professional treatment capabilities of large hospitals. In particular, there is a lack of empirical research on the role of PHC. After touring 5 cities in China, the WHO provided recommendations that were predominantly focused on secondary care and epidemiological tracking and control, and the role of PHC was missed again [14].

Distinguishing itself from secondary care for specialist treatment, PHC is regarded as the most inclusive, effective, and efficient approach to enhancing people's physical and mental health. PHC has great value in a strong, coordinated response to a public health crisis [10,15]. Recent studies show that a strong PHC foundation could effectively mitigate an epidemic. One such case is that of Singapore, which promptly instituted aggressive containment measures by establishing public health preparedness clinics that were supported in a sustained manner by the PHC network [16]. In contrast, PHC resources in the African Union are exceedingly scarce, which resulted in insufficient engagement when dealing with COVID-19 [17-20]. Even in countries with adequate PHC resources, such as the United Kingdom, the health system did not respond quickly and struggled to meet medical demands under a large-scale epidemic [21]. Legido-Quigley and colleagues [7] argued that well-developed integration was a key factor of services influencing resilience during the COVID-19 pandemic in high-performing health systems. Prompt communication and coordination among PHC, public health, and secondary care are essential [22].

At the end of 2022, COVID-19–related restrictions were lifted in China, and an epidemic wave caused by the highly transmissible Omicron SARS-CoV-2 variant placed health services in the country under extreme pressure, especially in metropolises. In Shanghai, which is the most populous metropolis in China and has a permanent population of 25 million, it is extremely difficult to deal with the spread of epidemic infections. At the end of 2022, Shanghai adopted an expansion strategy that involved allocating medical resources in secondary care institutions in a manner that favored patients with SARS-CoV-2 infection. At the same time, Shanghai, as a pilot city, was one of the first cities to promote a hierarchical diagnosis and treatment system based on PHC. As such, Shanghai provides an extremely rare opportunity to explore how the health system of a metropolis can actively respond to large-scale infections, as well as the key role of PHC in this system. In this study, we simulated the large-scale infections that occurred in Shanghai at the end of 2022 by using a simulated environment, wherein we reproduced Shanghai's response to the challenges of the fast-spreading epidemic. We then tested an alternative strategy that used a PHC-based integrated health system.

Methods

Ethical Considerations

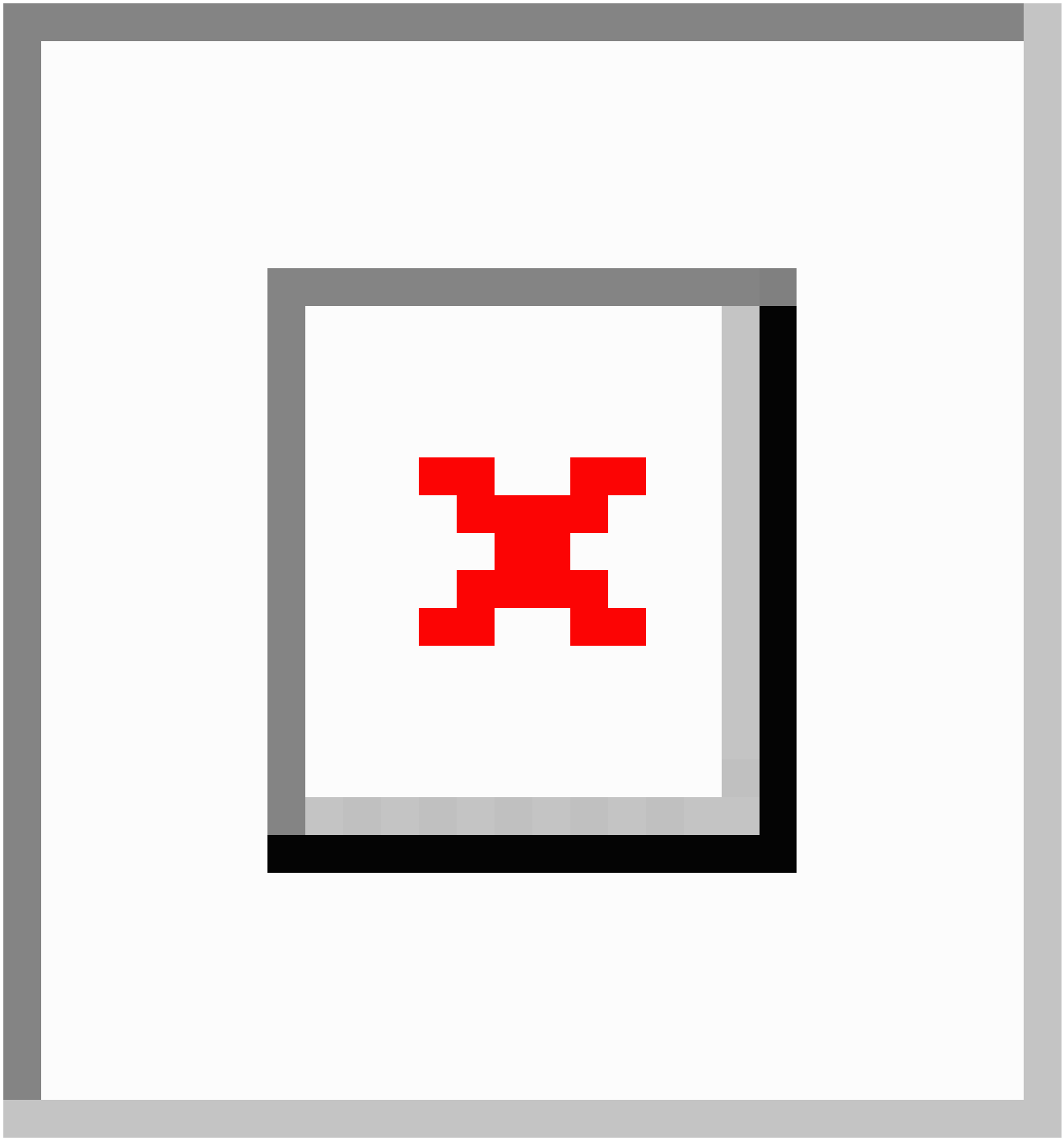
Ethics board review was not required, as this study only involved modeling and simulations. All modeling data came from public sources or published papers and did not involve ethical issues.

Study Design

System dynamics was applied in this study to simulate the mass infections and the health system performance in Shanghai. System dynamics models are established based on the feedback structures (loops) centered around the issue of concern [23,24]. The nonlinear dynamic behaviors derived from these feedback loops shed light on the underlying mechanisms that generate problematic system behaviors, which helps with understanding complex systems and finding fundamental solutions [25]. The use of system dynamics models is a suitable method for investigating public health issues that feature high-complexity systems [26,27]. In recent years, system dynamics has been widely used to model issues related to COVID-19 [28-31].

We developed a system dynamics–based model to replicate the health system in Shanghai after COVID-19–related policies were lifted. The following indicators of an overwhelmed health system were used: physician availability (the percentage of patients arriving at the hospital who could be treated) and bed availability (the percentage of patients needing hospitalization who could be admitted) in secondary hospitals. Shanghai's response to the soaring medical demands was to reallocate medical resources from other divisions to increase the supply of hospital physicians and beds for patients with COVID-19. This policy increased the capacity of secondary hospitals such that more patients could be treated and hospitalized. We also used the system dynamics model to establish a PHC-based integrated health system as an alternative option for addressing hospital overwhelm. The following three critical policy interventions were tested: first contact in PHC, identification of high-risk patients and recommendation to secondary care hospitals, and referral for a return to PHC for follow-up and recovery at the community level (Figure 1). Telemedicine services were also considered in PHC, with which more first contacts could be handled and the capacity of PHC to handle patients could be increased.

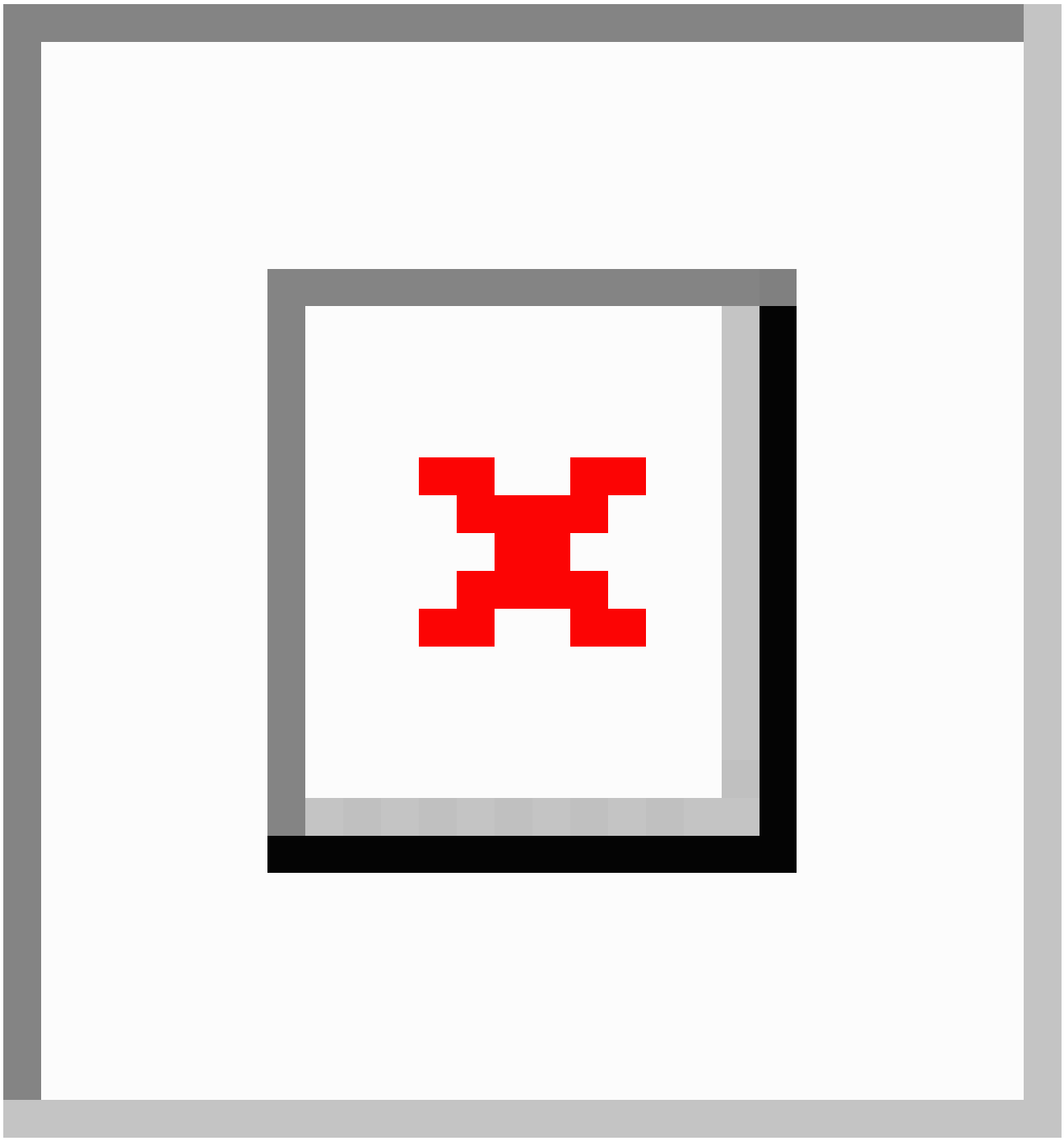
Figure 1. PHC-based health system. PHC: primary health care.



Model Structure

The Shanghai model includes the following two parts: the epidemic dynamics representing the mass infections that occurred when COVID-19 restrictions were lifted in Shanghai and the health care system response, as shown in Figure 2.

Figure 2. Shanghai system dynamics model of a PHC-based integrated system. *A*: infected population without symptoms; D: death; *E*: infected population during the incubation period; GB: getting better; GW: getting worse; HR: home recovery; *I*: infected population with symptoms; PHC: primary health care; R: recover; *RA*: population in *A* that has recovered; *S*: susceptible population without vaccination; *SV*: susceptible population with vaccination.



In Figure 2, the left part depicts an extension of the traditional Susceptible-Exposed-Infectious-Removed model, which we used to model the spread of COVID-19 in Shanghai and compute the number of symptomatic cases, of which a large proportion would require medical services. We disaggregated the total population into the following six groups.

SV and *S* represent the susceptible population with vaccination and the susceptible population without vaccination, respectively. The transformation of *SV* to *S* represents the waning effectiveness of COVID-19 vaccines, where ω is the waning effect of vaccination.

E represents the infected population during the incubation period. The transformation of *SV* to *E* and *S* to *E* represents the spread of the virus, where c is the contact rate, β is the transmission probability, and θ_1 is the effectiveness rate of vaccination against infection.

I and *A* represent the infected population with symptoms and the infected population without symptoms, respectively. α is the percentage of asymptomatic cases, and τ is the incubation period.

RA represents the population in *A* that has recovered. γ_1 is the recovery fraction among asymptomatic cases.

Patients with symptoms (ie, those from population *I*) link the left and right parts of the model. Some patients will recover at home, and others will visit a physician. Among those visiting a physician, some will first contact a PHC institution, while others will contact a secondary hospital directly. PHC institutions and secondary hospitals each have a specific capacity, and when this capacity is reached, new, excess patients cannot be treated and will have to return home. With regard to patients treated in PHC institutions and secondary hospitals, those with mild symptoms will be given prescriptions and sent home to recover. With regard to patients needing further treatment when presenting at the PHC level, general practitioners will recommend them to a secondary hospital; some patients will be hospitalized and become inpatients if hospital beds are available. Over time, some inpatients will recover, whereas others will develop severe illness and eventually recover or die. Patients who recover at home will either improve or worsen, as will untreated and treated patients from PHC institutions and secondary hospitals. The proportion of patients whose condition worsens is highest for untreated patients and lowest for treated patients. Some recovering inpatients in secondary care hospitals might recover at the community level if PHC can provide follow-up health management services. The model equations and parameter settings are detailed in sections 1 and 2 in Multimedia Appendix 1.

Data Source and Model Validation

We previously developed and validated a model of reopening in Shanghai that accounts for the epidemiological dynamics of the first Omicron wave in this metropolis during the first half of 2022 [32]. The model we established in this study was based on that previous model and was used to simulate the second Omicron wave, specifically the time when most intervention prevention control measures were lifted at the end of 2022. Data related to the spread of Omicron in Shanghai, such as the contact rate, transmission possibility, asymptomatic rate, incubation period, and recovery fraction, were obtained from previous literature about COVID-19 and, especially, Omicron (further details are reported in section 2 and Table S1 in Multimedia Appendix 1). Data related to individuals' behaviors, such as the rate of first contact in PHC and the rate of recovery at home, were set according to estimations based on our investigation of PHC, hospitals, and the community. Sensitivity tests for these parameters were conducted to check the robustness of the model (section 3.2 in Multimedia Appendix 1).

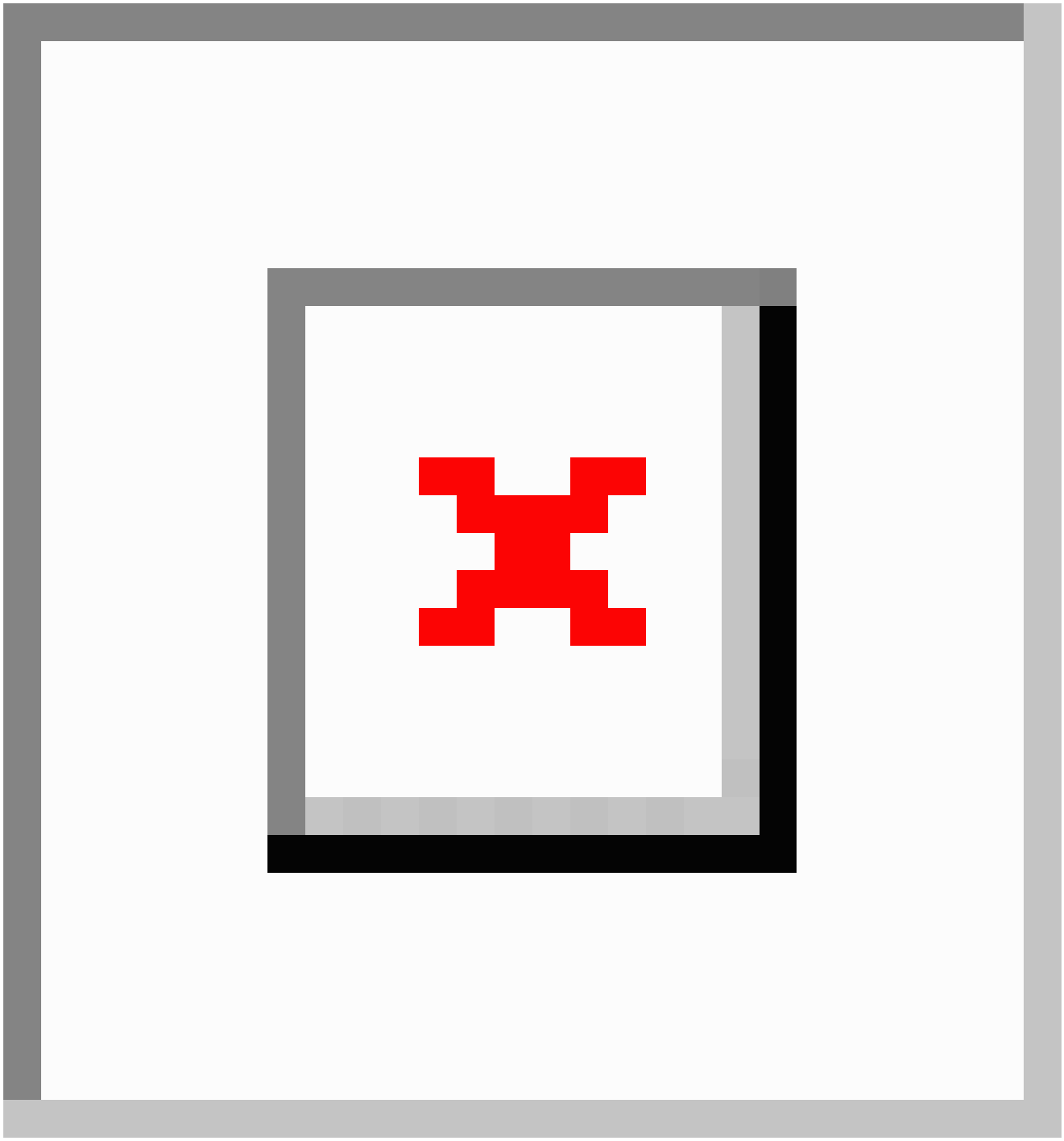
The validation of model behavior usually involves the comparison of simulation results with real-world data. Because mass COVID-19 testing was no longer required, accurate infection data were no longer available; news reports on viral infections and the level of pressure on medical resources were used as references. The model results revealed patterns that were similar to the actual situation during the second Omicron wave in Shanghai, thereby confirming the validity of the model (further details are reported in section 3 in Multimedia Appendix 1). As a result, we determined that this model could provide a simulated environment to facilitate the exploration of effective policies regarding the response to mass infections in a metropolis.

Results

Scenario 1: Medical Resource Reallocation in Secondary Care

When the strict intervention prevention controls were lifted, the policy focus changed from preventing the spread of COVID-19 to the timely treatment of patients with COVID-19. When facing massive increases in infections, physicians' availability could decline to as low as 55% if no interventions were adopted. In the case of Shanghai, a series of measures was taken to deal with the impact of large-scale infections on hospitals. When hospital physician capacity and hospital bed capacity were increased by 70% of the original capacities, the lowest physician availability and bed availability changed to approximately 85% and 70%, respectively. Moreover, bed shortages lasted approximately 8 days, which was around one-third of the bed shortage time for the scenario with no capacity extension. The peak number of severe cases decreased, as more patients could be treated promptly. Consequently, the cumulative number of deaths decreased to less than half of that for the scenario without additional resources, as shown in Figure 3.

Figure 3. Expanding capacity to meet soaring demand. Base: baseline; Ext cap 30: 30% capacity extension; Ext cap 50: 50% capacity extension; Ext cap 70: 70% capacity extension.

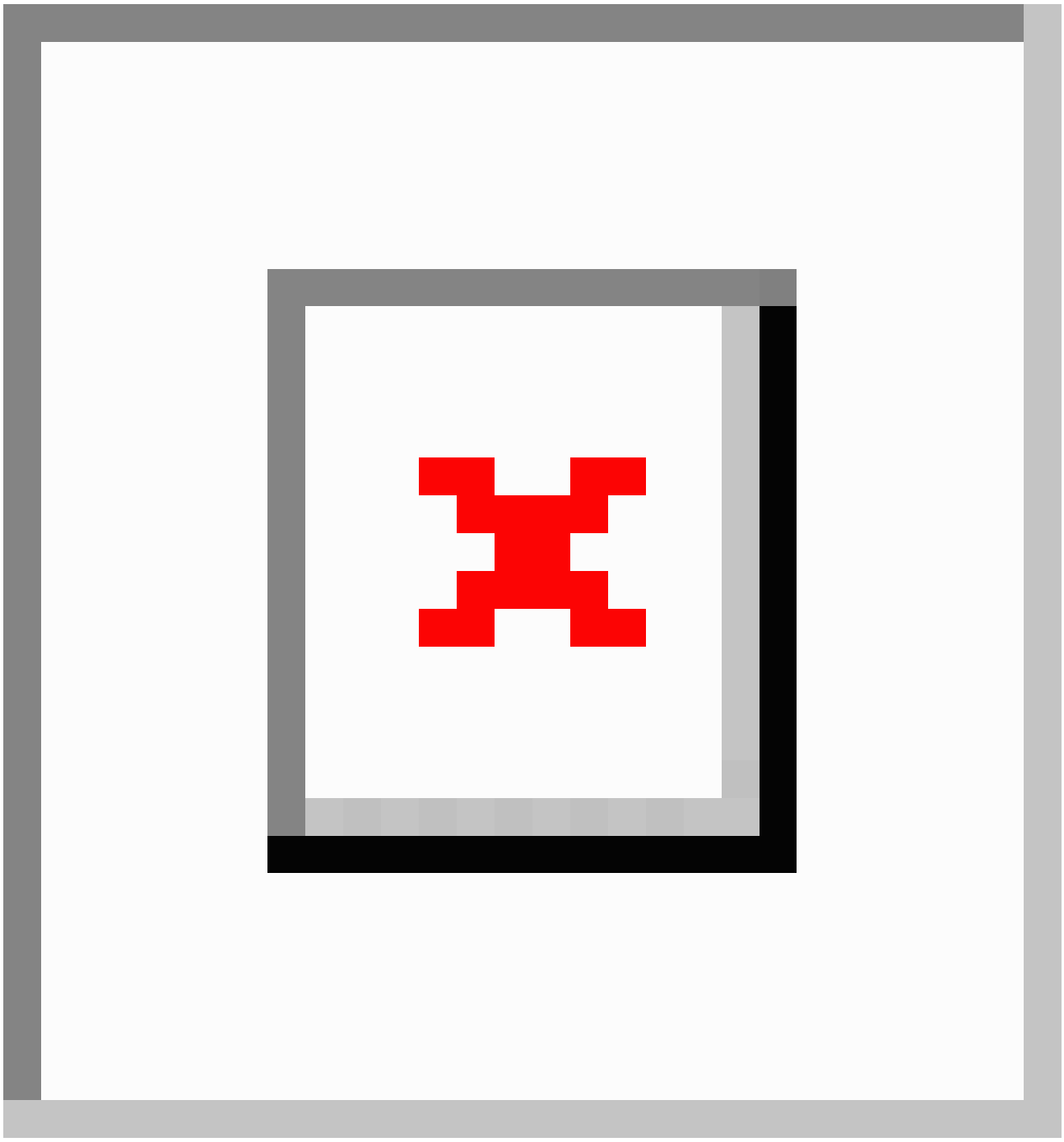


Scenario 2: PHC-Based Integrated Health Care System

Scenario 2.1: Increasing the Rate of First Contact in PHC Plus PHC Telemedicine

Facing huge increases in the number of patients, we examined ways to increase the rate of first contact in PHC, in which PHC telemedicine services were also considered. Under such circumstances, 6 scenarios were simulated, with the rate of first contact in PHC with and without telemedicine services set to 40%, 50%, and 60%. Figure 4 shows that replacing the worst scenario (40% rate of first contact in PHC without telemedicine) with the best scenario (60% rate of first contact in PHC with telemedicine) increased the lowest level of secondary hospital physician availability and that of secondary hospital bed availability by 32% (from 51% to 67%) and 111% (from 9% to 19%), respectively. Moreover, the duration of bed shortages dropped from approximately 30 days to approximately 20 days. Because more patients were promptly treated in the best scenario, the number of cumulative deaths decreased from 24,740 in the worst scenario to 15,837—a 56% decrease.

Figure 4. Scenarios with various rates of first contacts in primary health care with and without telemedicine. PHC FC 40: 40% rate of first contact in primary health care without telemedicine; PHC FC 50: 50% rate of first contact in primary health care without telemedicine; PHC FC 60: 60% rate of first contact in primary health care without telemedicine; PHC FC 40 + Telem 2: 40% rate of first contact in primary health care with telemedicine; PHC FC 50 + Telem 2: 50% rate of first contact in primary health care with telemedicine; PHC FC 60 + Telem 2: 60% rate of first contact in primary health care with telemedicine.

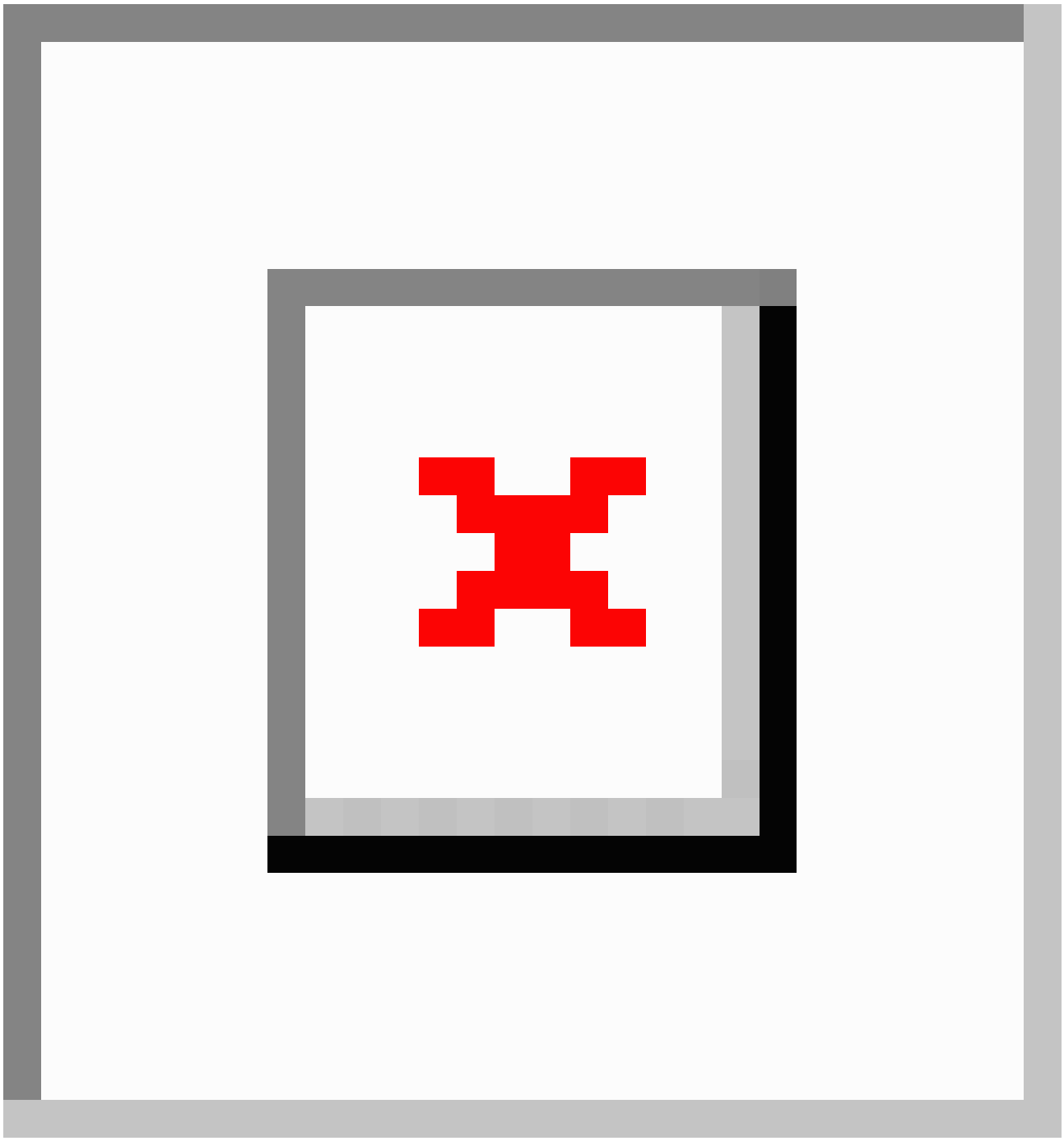


Scenario 2.2: Referral Recommendation Rate for High-Risk Patients

Two of the main tasks of general practitioners in PHC are to identify high-risk patients and provide referral recommendations to secondary care, which are related to the professional capabilities of general practitioners. We simulated the following four scenarios: (1) a recommendation rate of 80%, meaning that 20% of patients requiring advanced treatment in a hospital were not identified (ie, underrecommendation); (2) a recommendation rate of 100% without underrecommendation or overrecommendation, which is an ideal scenario; (3) a recommendation rate of 120%, meaning that 20% of patients were overreferred to secondary care; and (4) a recommendation rate of 100% but with underrecommendation and overrecommendation happening at the same time. Underrecommendation and overrecommendation, respectively, slightly increased and decreased the physician availability and bed availability. However, with underrecommendation, in which 20% of patients who needed to be treated in a hospital were not referred, some patients developed severe illness due to improper

treatment, leading to more severe cases and more cumulative deaths, as shown in Figure 5.

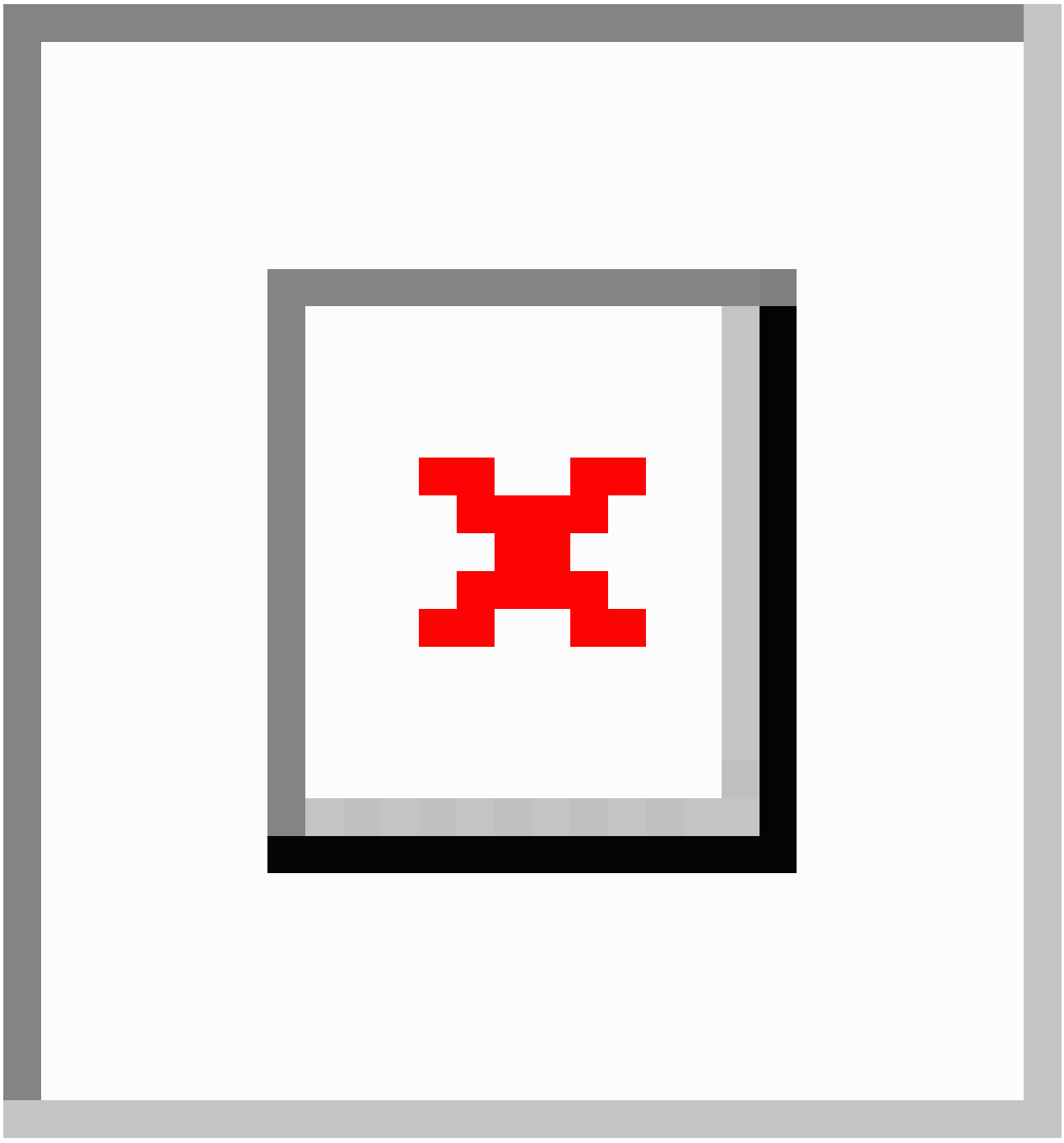
Figure 5. Scenarios involving general practitioners in primary health care with differing capabilities for identifying high-risk patients. “Ideal” refers to no underrecommendation and no overrecommendation. “Nonideal” refers to underrecommendation and overrecommendation happening at the same time.



Scenario 2.3: Follow-Up Health Management Services in PHC

When facing bed shortages, some patients with mild symptoms who are nearly recovered can be transferred to recover in the community if PHC can provide follow-up health management services. We simulated four scenarios under which 5%, 10%, 15%, and 20% of hospital inpatients were transferred back to PHC to recover. As shown in Figure 6, physician availability was not affected, but hospital bed availability improved. When 20% of inpatients could recover in the community, the lowest bed availability level reached approximately 20%, whereas this level reached 7% in the scenario where only 5% of inpatients were referred to PHC to recover. Moreover, the number of days with bed shortages was nearly halved. As a result, the peak number of severe cases decreased from 16,897 to 14,737, and the cumulative number of deaths dropped from 28,008 to 14,344.

Figure 6. Scenarios with different percentages of patients returning to primary health care for recovery. Health mngt 05: 5% rate of health management in the community; Health mngt 10: 10% rate of health management in the community; Health mngt 15: 15% rate of health management in the community; Health mngt 20: 20% rate of health management in the community.

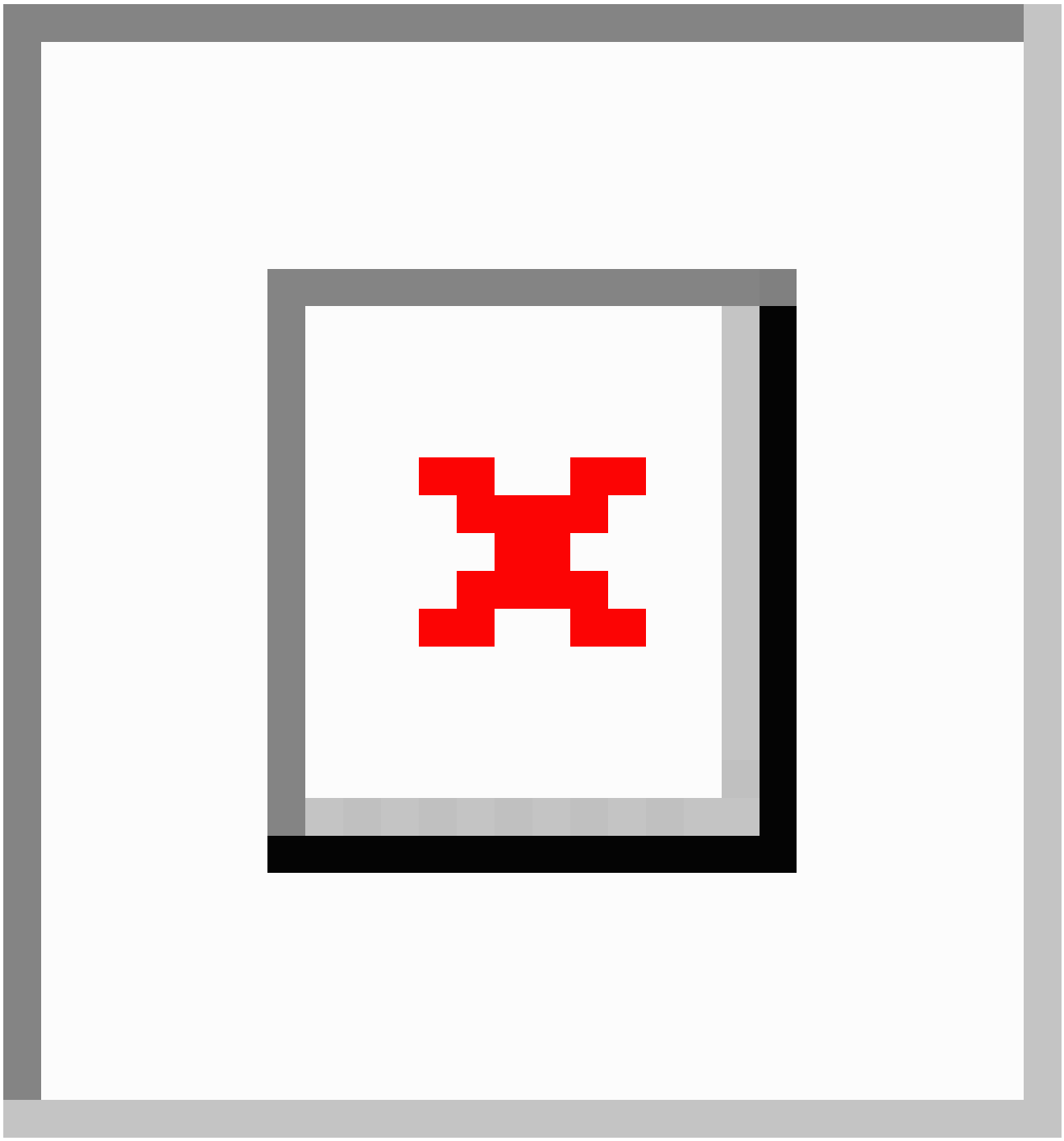


Scenario 2.4: Mixed Policy Interventions

We integrated the above three policy interventions to investigate the overall impact of a PHC-based system, as shown in Figure 7. The model simulation results showed that the lowest level of hospital physician availability ranged from 51% in the worst case (40% rate of first contact in PHC without telemedicine) to 69% in the best case (60% rate of first contact in PHC with telemedicine). The lowest level of hospital bed availability varied even more, ranging from 6% in the worst case (40% rate of first contact in PHC without telemedicine and 5% rate of health management in the community) to 51% in the best case (60% rate of first contact in PHC with telemedicine, recommendation rate of 80%, and 20% rate of health management in the community). With regard to the number of severe cases, the worst case peaked at 20,876 severe cases (40% rate of first contact in PHC without telemedicine, underrecommendation rate of 20%, and 5% rate of health management in the community), and the best case peaked at 11,984 severe cases (60% rate of first contact in PHC with telemedicine, recommendation rate of 100%, and 20% rate of health management in the community)—a decrease of approximately 75%. As for cumulative deaths, the worst case was 37,369 deaths, and the best case was only 7946 deaths—a 79% reduction. Furthermore, we identified the following relatively optimal policy intervention mix: a 60% rate of first

contact in PHC with telemedicine services, a 20% rate of referral to return to PHC, and a recommendation rate of 100% to 120%.

Figure 7. Overall impact of integrated primary health care (PHC). The red, orange, green, and blue areas represent PHC with or without telemedicine, the rate of FC in PHC, the rate of PHC recommendation to secondary care, and the rate of referral to return to PHC, respectively. The gray areas represent the biggest medical resource gaps, including hospital physician availability and hospital bed availability, and the resulting peak number of SCs and cumulative deaths. FC: first contact; SC: severe case.



Discussion

Principal Findings

The experience of hospital overwhelm during the COVID-19 pandemic highlights the need to reflect on existing health systems and search for more proactive solutions during an epidemic. In this study, we constructed a simulated policy environment to replicate Shanghai's response to the mass SARS-CoV-2 infections that occurred once restrictions were lifted. Specifically, the Shanghai Municipal Health Commission deployed medical resources in secondary care hospitals, in a manner that favored patients with SARS-CoV-2 infection, as quickly as possible. This strategy included increasing the availability of beds and reallocating more medical staff from other departments to promptly treat critically ill patients and prevent deaths. This efficient and decisive response allowed Shanghai to avoid the large-scale congestion and overwhelm of the health care system. However, Shanghai's strategy worked under the assumptions that advanced medical resources would

be available and could be deployed, and the strategy largely depended on the government's strong decision-making and coordination capabilities. At the same time, we realize that relying on PHC to alleviate congestion is an important strategy to achieve the effective allocation of medical resources, rather than only relying on temporary expansion in secondary care [33].

We proposed an alternative PHC-based strategy and tested this in a simulated policy simulation environment. We tested the rate of first contact in PHC, the rate of identifying high-risk patients for recommendation to a specialist, and the rate of return to PHC for recovery. According to the simulation results, increasing the rate of first contact in PHC could effectively alleviate the shortage of specialists in large hospitals. Additionally, telemedicine application in PHC contributed substantially to reducing congestion within hospitals engaged in COVID-19 treatment. In our model, a 60% rate of first contact in PHC with telemedicine could increase the lowest level of secondary hospital physician availability from 51% to 67% and reduce the number of cumulative deaths by 9630. The value of first contact in PHC for patients is receiving immediate medical treatment to avoid severe illness or death caused by delays in treatment, as well as reducing the shortage of medical resources in secondary hospitals. COVID-19 has accelerated the development of telemedicine. Alexander and colleagues [34] used a nationally representative audit of outpatient care to characterize primary care delivery in the United States and found that the pandemic was associated with a >25% decrease in primary care volume, which has been offset in part by increases in the delivery of telemedicine. Some believe that the boom in telemedicine during the COVID-19 pandemic could worsen health disparities [35], especially for racial and ethnic minority groups; those living in rural areas; and individuals with limited English proficiency, low literacy, or low income [36]. Nevertheless, telemedicine is an inevitable future developmental trend.

The rate of identifying high-risk patients is a crucial indicator of PHC worker capacity. We found that underidentification could result in more severe illness and more deaths, whereas overidentification could increase congestion in hospitals to some degree. For example, in the scenario with a 50% rate of first contact in PHC with telemedicine, a 120% recommendation rate reduced hospital specialist availability from 61% to 60%, whereas an 80% recommendation rate increased hospital specialist availability from 61% to 63%. A similar impact was observed on hospital bed availability. However, underrecommendation resulted in some patients (ie, those needing further treatment) failing to seek timely medical care and thus higher rates of severe illness and an increase of 3265 cumulative deaths. According to the simulation results, the effect of accurately identifying high-risk patients is limited in the existing system, possibly due to a low rate of first contact in PHC. Unlike countries in Europe and North America, China has a loose medical referral system rather than a strict referral system based on first contact in PHC [37]. China established its PHC system after the new health care reform in 2009 [38]. In October 2016, the Chinese government launched the Healthy China 2030 initiative, in which a critical component is developing a patient referral model [39]. In contrast, gatekeeping systems can ensure the efficient use of scarce medical resources in secondary care; to date, there has been no action plan to enforce the patient referral model [37]. The rate of first contact in PHC has remained at approximately 30% to 50% for the past 10 years. However, in scenarios where PHC first contact–based referral is strictly implemented, such as in the United Kingdom [40], we believe that accurate risk identification in PHC is important.

We also considered the rate of return to PHC for recovery in the community, which can accelerate bed turnover in secondary hospitals. The model simulation results showed that increasing the rate of return to PHC from 5% to 20% would increase bed availability from 6% to 16%, thereby reducing the number of cumulative deaths by approximately 13,000. According to the WHO, referral is a bidirectional process that acknowledges not only the role of the specialist but also the critical role of PHC workers in coordinating patient care over the longer term [41]. In May 2023, Shanghai issued an important document—*Implementation Plan to Further Enhance the City's Community Health Service Capabilities*—focusing on strengthening 4 functions in community health centers—PHC, health management, rehabilitation, and nursing—as well as the primary public health network [42]. COVID-19 has definitely brought challenges to PHC, but it has also provided new opportunities.

Interestingly, we also found a multiplier effect with combined policy interventions. For example, offering telemedicine services, increasing the rate of first contact in PHC from 40% to 60%, and raising the rate of referral to return to PHC from 5% to 20% respectively increased bed availability by 16.67%, 50%, and 167%. However, when combined, these policies increased the lowest bed availability level by 683%. Optimal policy intervention combinations are widely applied in health, climate change, and economics (eg, funding instruments). Policy mixing implies a focus on trade-off interactions and interdependencies among different policies, as they affect the extent to which the intended policy outcomes are achieved. It provides a window of opportunity to reconsider basic and often hidden assumptions to better deal with complex, multilevel, multiactor realities [43]. In this study, we identified a relatively optimal policy combination (ie, a 60% rate of first contact in PHC, a 110% recommendation rate, and a 20% rate of return to PHC) that could establish a strong PHC foundation and increase health system resilience by reducing medical resource gaps in responding to public health emergencies. The interplay of policies and instruments, as well as the deliberate design of policy mixes and portfolios of interventions, has received surprisingly little practical and theoretical attention so far and is vastly underrated [44].

Using the scenario of reopening in Shanghai, we built a health care system for metropolises to deal with large-scale infections and verified the role of PHC through a system simulation model. However, our study has some limitations. First, real-world data were missed, especially epidemiological data, as mass COVID-19 testing was canceled. We validated our model based on information from news reports indicating the development of the Omicron wave and web-based information. Second, data

related to individuals' behaviors, such as the rate of first contact in PHC and the rate of recovery at home, are not available. We estimated these parameters according to our investigation of PHC institutions, hospitals, and communities. Third, this study simulated a PHC-based integrated health system responding to large-scale infections (including parameters such as first-contact rate, referral rate, and recovery fraction), but our models did not tell us how the integrated system could be improved. More attention should be paid to integrated health systems in the near future by conducting more case studies or implementation research.

Conclusions

Rather than focusing on secondary care, in this study, we proposed an alternative—strengthening the health system via a bottom-up approach by using PHC as a foundation to better respond to a public health emergency. Per our PHC-based health system, an optimal PHC-based integrated strategy would be to have a 60% rate of first contact in PHC, a 110% recommendation rate, and a 20% rate of return to PHC, which could increase health system resilience during public health emergencies. A robust PHC-based integrated health system, in addition to the temporary deployment of medical resources in secondary care, could maximize the use of limited medical resources to actively respond to large-scale increases in infections. This study provides an optimal solution for constructing a PHC-based integrated health system to respond to large-scale infections. We acknowledge that there is a long way to go to achieve an integrated health system, as getting PHC to communicate and interact seamlessly with secondary care is extremely challenging globally. We advocate increasing investments in PHC to promote its overall development and conducting more research on integrated health systems in the near future.

Acknowledgments

This study was funded by the Three-Year Action Program of Shanghai Municipality for Strengthening the Construction of Public Health System (grant GWVI-11.2-YQ54), Science and Technology Committee of Shanghai Municipality Soft Science Research Plans (grant 23692113400), the National Natural Science Foundation of China (grant 72274122), and the National Social Science Foundation of China (grant 22BGL240).

Data Availability

The data used in this study are publicly available on the National Health Commission of the People's Republic of China website [45]. Our model code is available from the corresponding author on request.

Authors' Contributions

JH, HL, and YQ conceptualized this study. JH and YQ wrote the original draft and reviewed and edited the manuscript. YY was responsible for data visualization. LZ was responsible for data collection and data analysis. YQ was responsible for the methodology.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Model description and definitions, parameter settings, and model validation.
[DOCX File, 691 KB - medinform_v12i1e54355_app1.docx]

References

1. WHO COVID-19 dashboard. World Health Organization. URL: https://covid19.who.int/ [accessed 2023-07-28]
2. Armocida B, Formenti B, Ussai S, Palestra F, Missoni E. The Italian health system and the COVID-19 challenge. Lancet Public Health 2020 May;5(5):e253. [doi: 10.1016/S2468-2667(20)30074-8] [Medline: 32220653]
3. da Silva SJR, Pena L. Collapse of the public health system and the emergence of new variants during the second wave of the COVID-19 pandemic in Brazil. One Health 2021 Dec;13:100287. [doi: 10.1016/j.onehlt.2021.100287] [Medline: 34222607]
4. El Bcheraoui C, Weishaar H, Pozo-Martin F, Hanefeld J. Assessing COVID-19 through the lens of health systems' preparedness: time for a change. Global Health 2020 Nov 19;16(1):112. [doi: 10.1186/s12992-020-00645-5] [Medline: 33213482]
5. Han E, Tan MMJ, Turk E, et al. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. Lancet 2020 Nov 7;396(10261):1525-1534. [doi: 10.1016/S0140-6736(20)32007-9] [Medline: 32979936]
6. Mahendradhata Y, Andayani NLPE, Hasri ET, et al. The capacity of the Indonesian healthcare system to respond to COVID-19. Front Public Health 2021 Jul 7;9:649819. [doi: 10.3389/fpubh.2021.649819] [Medline: 34307272]
7. Legido-Quigley H, Asgari N, Teo YY, et al. Are high-performing health systems resilient against the COVID-19 epidemic? Lancet 2020 Mar 14;395(10227):848-850. [doi: 10.1016/S0140-6736(20)30551-1] [Medline: 32151326]

8. Tangcharoensathien V, Bassett MT, Meng Q, Mills A. Are overwhelmed health systems an inevitable consequence of COVID-19? experiences from China, Thailand, and New York State. BMJ 2021 Jan 22;372:n83. [doi: 10.1136/bmj.n83] [Medline: 33483336]
9. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Health Surveill 2020 Apr 2;6(2):e18810. [doi: 10.2196/18810] [Medline: 32238336]
10. Lim WH, Wong WM. COVID-19: notes from the front line, Singapore's primary health care perspective. Ann Fam Med 2020 May;18(3):259-261. [doi: 10.1370/afm.2539] [Medline: 32393562]
11. Lauriola P, Martín-Olmedo P, Leonardi GS, et al. On the importance of primary and community healthcare in relation to global health and environmental threats: lessons from the COVID-19 crisis. BMJ Glob Health 2021 Mar;6(3):e004111. [doi: 10.1136/bmjgh-2020-004111] [Medline: 33692145]
12. Malaysia: a primary health care case study in the context of the COVID-19 pandemic. World Health Organization. 2023 Aug 28. URL: https://www.who.int/publications/i/item/9789240076723 [accessed 2024-02-02]
13. Frieden TR, Lee CT, Lamorde M, Nielsen M, McClelland A, Tangcharoensathien V. The road to achieving epidemic-ready primary health care. Lancet Public Health 2023 May;8(5):e383-e390. [doi: 10.1016/S2468-2667(23)00060-9] [Medline: 37120262]
14. Kupferschmidt K, Cohen J. Can China's COVID-19 strategy work elsewhere? Science 2020 Mar 6;367(6482):1061-1062. [doi: 10.1126/science.367.6482.1061] [Medline: 32139521]
15. Declaration of Astana. World Health Organization. 2018. URL: https://www.who.int/docs/default-source/primary-health/declaration/gcphc-declaration.pdf [accessed 2023-07-28]
16. Wong SYS, Tan DHY, Zhang Y, et al. A tale of 3 Asian cities: how is primary care responding to COVID-19 in Hong Kong, Singapore, and Beijing. Ann Fam Med 2021;19(1):48-54. [doi: 10.1370/afm.2635] [Medline: 33431392]
17. Kavanagh MM, Erondu NA, Tomori O, et al. Access to lifesaving medical resources for African countries: COVID-19 testing and response, ethics, and politics. Lancet 2020 May 30;395(10238):1735-1738. [doi: 10.1016/S0140-6736(20)31093-X] [Medline: 32386564]
18. Universal health coverage (UHC) in Africa: a framework for action: main report (English). : World Bank Group; 2016 URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/735071472096342073/main-report [accessed 2023-07-28]
19. Wood G. Think 168,000 ventilators is too few? Try three. The Atlantic. 2020 Apr 10. URL: https://www.theatlantic.com/ideas/archive/2020/04/why-covid-might-hit-african-nations-hardest/609760/ [accessed 2023-07-28]
20. Huang P. 10,000 cases and 500 deaths in Africa. Health officials say it's just the beginning. NPR. 2020 Apr 8. URL: https://www.npr.org/sections/coronavirus-live-updates/2020/04/08/830209940/10-000-cases-and-500-deaths-in-africa-health-officials-say-its-just-the-beginnin [accessed 2023-07-28]
21. Lal A, Erondu NA, Heymann DL, Gitahi G, Yates R. Fragmented health systems in COVID-19: rectifying the misalignment between global health security and universal health coverage. Lancet 2021 Jan 2;397(10268):61-67. [doi: 10.1016/S0140-6736(20)32228-5] [Medline: 33275906]
22. Park S, Elliott J, Berlin A, Hamer-Hunt J, Haines A. Strengthening the UK primary care response to COVID-19. BMJ 2020 Sep 25;370:m3691. [doi: 10.1136/bmj.m3691] [Medline: 32978177]
23. Forrester JW. Industrial Dynamics: MIT Press; 1961.
24. Sterman JD. Business Dynamics: Systems Thinking and Modeling for a Complex World: McGraw-Hill Education; 2000.
25. Sterman JD. Learning from evidence in a complex world. Am J Public Health 2006 Mar;96(3):505-514. [doi: 10.2105/AJPH.2005.066043] [Medline: 16449579]
26. Homer JB, Hirsch GB. System dynamics modeling for public health: background and opportunities. Am J Public Health 2006 Mar;96(3):452-458. [doi: 10.2105/AJPH.2005.062059] [Medline: 16449591]
27. Darabi N, Hosseinichimeh N. System dynamics modeling in health and medicine: a systematic literature review. Syst Dyn Rev 2020 Mar 22;36(1):29-73. [doi: 10.1002/sdr.1646]
28. Rahmandad H, Sterman J. Quantifying the COVID-19 endgame: is a new normal within reach? Syst Dyn Rev 2022 Aug 24. [doi: 10.1002/sdr.1715] [Medline: 36246868]
29. Qian Y, Xie W, Zhao J, et al. Investigating the effectiveness of re-opening policies before vaccination during a pandemic: SD modelling research based on COVID-19 in Wuhan. BMC Public Health 2021 Sep 7;21(1):1638. [doi: 10.1186/s12889-021-11631-w] [Medline: 34493226]
30. Zhao J, Jia J, Qian Y, Zhong L, Wang J, Cai Y. COVID-19 in Shanghai: IPC policy exploration in support of work resumption through system dynamics modeling. Risk Manag Healthc Policy 2020 Oct 8;13:1951-1963. [doi: 10.2147/RMHP.S265992] [Medline: 33116976]
31. Huang J, Qian Y, Shen W, et al. Optimizing national border reopening policies in the COVID-19 pandemic: a modeling study. Front Public Health 2022 Nov 30;10:979156. [doi: 10.3389/fpubh.2022.979156] [Medline: 36530669]
32. Qian Y, Cao S, Zhao L, Yan Y, Huang J. Policy choices for Shanghai responding to challenges of Omicron. Front Public Health 2022 Aug 9;10:927387. [doi: 10.3389/fpubh.2022.927387] [Medline: 36016887]

33. Huang J, Liu Y, Zhang T, et al. Can family doctor contracted services facilitate orderly visits in the referral system? a frontier policy study from Shanghai, China. Int J Health Plann Manage 2022 Jan;37(1):403-416. [doi: 10.1002/hpm.3346] [Medline: 34628680]
34. Alexander GC, Tajanlangit M, Heyward J, Mansour O, Qato DM, Stafford RS. Use and content of primary care office-based vs telemedicine care visits during the COVID-19 pandemic in the US. JAMA Netw Open 2020 Oct 1;3(10):e2021476. [doi: 10.1001/jamanetworkopen.2020.21476] [Medline: 33006622]
35. Ortega G, Rodriguez JA, Maurer LR, et al. Telemedicine, COVID-19, and disparities: policy implications. Health Policy Technol 2020 Sep;9(3):368-371. [doi: 10.1016/j.hlpt.2020.08.001] [Medline: 32837888]
36. Patton-López MM. Communities in action: pathways to health equity. J Nutr Educ Behav 2022 Jan;54(1):P94-P95. [doi: 10.1016/j.jneb.2021.09.012]
37. Xiao Y, Chen X, Li Q, Jia P, Li L, Chen Z. Towards healthy China 2030: modeling health care accessibility with patient referral. Soc Sci Med 2021 May;276:113834. [doi: 10.1016/j.socscimed.2021.113834] [Medline: 33774532]
38. Tao W, Zeng Z, Dang H, et al. Towards universal health coverage: lessons from 10 years of healthcare reform in China. BMJ Glob Health 2020 Mar 19;5(3):e002086. [doi: 10.1136/bmjgh-2019-002086] [Medline: 32257400]
39. Yang J, Siri JG, Remais JV, et al. The Tsinghua-Lancet Commission on Healthy Cities in China: unlocking the power of cities for a healthy China. Lancet 2018 May 26;391(10135):2140-2184. [doi: 10.1016/S0140-6736(18)30486-0] [Medline: 29678340]
40. Forrest CB. Primary care in the United States: primary care gatekeeping and referrals: effective filter or failed experiment? BMJ 2003 Mar 29;326(7391):692-695. [doi: 10.1136/bmj.326.7391.692] [Medline: 12663407]
41. Hort K, Gilbert K, Basnayaka P, Annear PL. Strategies to strengthen referral from primary care to secondary care in low- and middle-income countries. World Health Organization. 2019. URL: https://iris.who.int/bitstream/handle/10665/325734/9789290227090-eng.pdf?sequence=1&isAllowed=y [accessed 2023-07-28]
42. Implementation plan to further enhance the city's community health service capabilities [Article in Chinese]. Shanghai Municipal People's Government. 2023 May 10. URL: https://www.shanghai.gov.cn/nw12344/20230510/55a194b734f54655ba5fc3bc60982a5d.html [accessed 2023-07-28]
43. Flanagan K, Uyarra E, Laranja M. Reconceptualising the 'policy mix' for innovation. Res Policy 2011 Jun;40(5):702-713. [doi: 10.1016/j.respol.2011.02.005]
44. Edler J, Cunningham P, Flanagan K, Laredo P. Innovation policy mix and instrument interaction: a review. : NESTA; 2013 URL: https://research.manchester.ac.uk/en/publications/innovation-policy-mix-and-instrument-interaction-a-review [accessed 2023-07-28]
45. 疫情通报 [Article in Chinese]. National Health Commission of the People's Republic of China. URL: http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml [accessed 2024-04-22]

Abbreviations

PHC: primary health care
WHO: World Health Organization

Edited by C Lovis; submitted 07.11.23; peer-reviewed by S Kreindler, W Tao; revised version received 04.03.24; accepted 10.03.24; published 03.06.24.

Please cite as:
Huang J, Qian Y, Yan Y, Liang H, Zhao L
Addressing Hospital Overwhelm During the COVID-19 Pandemic by Using a Primary Health Care–Based Integrated Health System: Modeling Study
JMIR Med Inform 2024;12:e54355
URL: https://medinform.jmir.org/2024/1/e54355
doi:10.2196/54355

Effect of Implementing an Informatization Case Management Model on the Management of Chronic Respiratory Diseases in a General Hospital: Retrospective Controlled Study

Yi-Zhen Xiao[1], MBBS; Xiao-Jia Chen[1], MBBS; Xiao-Ling Sun[1], MBBS; Huan Chen[1], MM; Yu-Xia Luo[1], MBBS; Yuan Chen[1], MBBS; Ye-Mei Liang[2], MM

1

2

Corresponding Author:
Ye-Mei Liang, MM

Abstract

Background: The use of chronic disease information systems in hospitals and communities plays a significant role in disease prevention, control, and monitoring. However, there are several limitations to these systems, including that the platforms are generally isolated, the patient health information and medical resources are not effectively integrated, and the "Internet Plus Healthcare" technology model is not implemented throughout the patient consultation process.

Objective: The aim of this study was to evaluate the efficiency of the application of a hospital case management information system in a general hospital in the context of chronic respiratory diseases as a model case.

Methods: A chronic disease management information system was developed for use in general hospitals based on internet technology, a chronic disease case management model, and an overall quality management model. Using this system, the case managers provided sophisticated inpatient, outpatient, and home medical services for patients with chronic respiratory diseases. Chronic respiratory disease case management quality indicators (number of managed cases, number of patients accepting routine follow-up services, follow-up visit rate, pulmonary function test rate, admission rate for acute exacerbations, chronic respiratory diseases knowledge awareness rate, and patient satisfaction) were evaluated before (2019 - 2020) and after (2021 - 2022) implementation of the chronic disease management information system.

Results: Before implementation of the chronic disease management information system, 1808 cases were managed in the general hospital, and an average of 603 (SD 137) people were provided with routine follow-up services. After use of the information system, 5868 cases were managed and 2056 (SD 211) patients were routinely followed-up, representing a significant increase of 3.2 and 3.4 times the respective values before use (U=342.779; P<.001). With respect to the quality of case management, compared to the indicators measured before use, the achievement rate of follow-up examination increased by 50.2%, the achievement rate of the pulmonary function test increased by 26.2%, the awareness rate of chronic respiratory disease knowledge increased by 20.1%, the retention rate increased by 16.3%, and the patient satisfaction rate increased by 9.6% (all P<.001), while the admission rate of acute exacerbation decreased by 42.4% (P<.001) after use of the chronic disease management information system.

Conclusions: Use of a chronic disease management information system improves the quality of chronic respiratory disease case management and reduces the admission rate of patients owing to acute exacerbations of their diseases.

(JMIR Med Inform 2024;12:e49978) doi:10.2196/49978

KEYWORDS

chronic disease management; chronic respiratory disease; hospital information system; informatization; information system; respiratory; pulmonary; breathing; implementation; care management; disease management; chronic obstructive pulmonary disease; case management

Introduction

Chronic obstructive pulmonary disease (COPD) and asthma are examples of common chronic respiratory diseases. The prevalence of COPD among people 40 years and older in China is estimated to be 13.7%, with the total number of patients reaching nearly 100 million. The lengthy disease cycle, recurrent acute exacerbations, and low control rate were found to have a significant impact on the prognosis and quality of life of middle-aged and older patients with COPD [1,2]. Therefore, to decrease the morbidity and disability rates and enhance the quality of life of all patients with chronic respiratory diseases, it is crucial to investigate effective prevention and treatment methods and establish a life cycle management model for chronic respiratory diseases.

Since the development of information technology, the internet and medical technology have been applied to the management of chronic diseases [3]. The chronic disease information systems adopted in hospitals and communities, along with mobile medical apps, can enhance the self-management capabilities of patients and play a significant role in disease prevention, control, and monitoring [4-9]. However, the existing platforms are generally isolated, the patient health information and medical resources are not effectively integrated, and the Internet Plus Healthcare technology model is not implemented throughout the patient consultation process [3,9].

Yulin First People's Hospital developed a chronic disease management information system based on the hospital information system (HIS) to fully and effectively utilize the medical resources in hospitals and to better support and adapt the system to the needs of patients with chronic diseases. In this study, we evaluated the impact of the use of this system on the efficacy of case management for patients with chronic respiratory diseases.

Methods

Chronic Respiratory Diseases Case Management Model Prior to Implementation of the Chronic Disease Management Information System

Yulin First People's Hospital is a public grade-3 general hospital with 2460 open beds, a specialty clinic in the Department of Pulmonary and Critical Care Medicine, and 180 beds in the Inpatient Department. Chronic respiratory diseases case management was initiated in 2019, which did not involve the use of an information system and was implemented by a chronic respiratory diseases case management team led by two nurses qualified as case managers, one chief physician, two supervisor nurses, and one technician. Under this system, patients with COPD, bronchial asthma, bronchiectasis, pulmonary thromboembolism, lung cancer, and lung nodules were managed using the traditional inpatient-outpatient-home chronic respiratory diseases case management model, including 1024 cases managed from 2019 to 2020. Except for medical prescriptions and electronic medical records, the patient case management information such as the basic information form, follow-up form, patient enrollment form, inpatient follow-up register, patient medication and inhalation device use records, smoking cessation and vaccination records, and pulmonary rehabilitation and health education records was managed using Microsoft Excel forms that were regularly printed for filing.

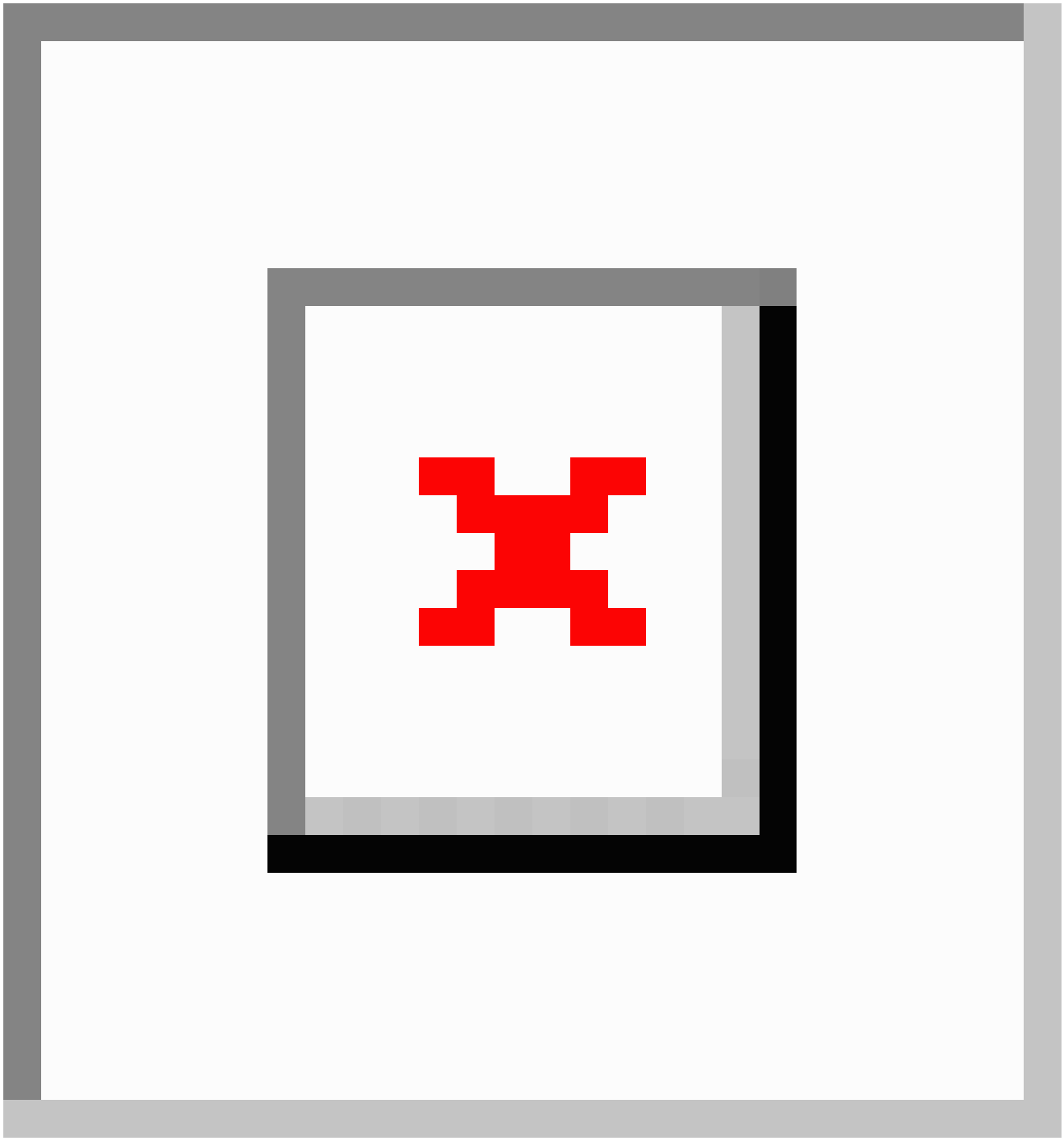
Establishment of a Management Information System for Chronic Diseases

The information carrier forming the basis of the management information system is constituted by the model of internet technology, chronic disease case management models, and overall quality management. The key technology is to establish a scientific, refined, and feasible follow-up pathway according to the methods and procedures of chronic disease case management based on the guidelines for the diagnosis and treatment of single chronic diseases. The closed-loop management of the clinical pathway was conducted in accordance with the Deming cycle (plan-do-check-act), and dynamic monitoring of single-disease health-sensitive and quality-sensitive indicators was carried out. The successfully developed system was installed on the hospital server to connect personal terminals (medical terminals and customer apps) to the existing HIS, which includes electronic medical records and medical advice.

Using the single-disease path assessment or plan scale as a framework, the system can automatically collect and integrate the majority of the medical information of patients with chronic respiratory diseases and provide these patients with inpatient, outpatient, and home intelligent medical services. Patients with chronic diseases who enroll in use of the system can use the app to schedule appointments for medical guidance, payment, and result queries; receive health guidance information; perform self-health assessments; write a treatment diary; and obtain medical communication materials.

The medical terminal consists of five functional modules: user entry, data statistics and query, quality control, knowledge base, and module management. As the core of the system, the user entry module can manage case information in seven steps: enrollment, assessment, planning, implementation, feedback, evaluation, and settlement [10-14]. Each step has a corresponding assessment record scale as well as the health-sensitive and quality-sensitive indicators. The structure of the HIS-based chronic disease management information system is shown in Figure 1.

Figure 1. Main structure of the hospital information system–based chronic disease management information system. *ICD*: *International Classification of Diseases*.



Implementation of the Chronic Disease Information Management System

Overview

Using the chronic disease management information system, two full-time case managers oversaw the case management of 2747 patients diagnosed with six diseases among chronic respiratory diseases between 2021 and 2022. The operation process was broken down into enrollment, assessment, planning, implementation, evaluation, feedback, and settlement stages.

Enrollment

Case managers entered the system through the medical app, selected a disease and an enrolled patient from the list of patients (the system automatically captures the patient's name and ID number according to the *International Classification of Diseases* [*ICD*] code) in accordance with the chronic respiratory diseases diagnostic criteria to sign the enrollment contract and determine the relationship between the personal information and data [15-19].

Assessment

The system can be seamlessly integrated with multiple workstations on the HIS to automatically capture the basic

information, electronic medical records, medical advice, and inspection materials, and can generate questionnaires or assessment scales for patients with chronic respiratory diseases such as the COPD Assessment Test, Asthma Control Test, modified Medical Research Council scale, form for lung function test results, inhalation device technique evaluation form, 6-minute walk test record, rehabilitation assessment form, health promotion form, and nutritional assessment form. The above materials can be added or removed based on the requirements for individual patients.

Planning

The case managers drafted the follow-up plan based on the patient assessment criteria and included the patients on the 1-, 3-, 6-, and 12-month follow-up lists. If the patient satisfied the self-management and indicator control requirements after follow-up, they could be settled and included in the annual follow-up cohort. Case managers can set up follow-up warning and treatment, involving the return visit plan, health education, follow-up content, pathway, and time, and notify the patients and nurses on day 7 and at months 1, 3, 6, and 12 after discharge. The nurses should promptly deal with patients who miss their scheduled follow-up visit.

Implementation

During the inpatient or outpatient care, supervising physicians, nurses, and patients collaborated with each other to implement the treatment. Case managers monitored the patients, evaluated them, documented the results, interpreted various test indicators, and provided health guidance. The chronic disease management information system acquired the corresponding data for chronic disease–sensitive indicators from outpatient and inpatient orders and medical records automatically. The chronic respiratory diseases management team reviewed the patients' conditions and the dynamics of chronic disease–sensitive indicators to make accurate decisions based on the current situation. The outpatient physicians obtained the single-disease package advice and personalized prescriptions to modify the diagnosis and treatment scheme.

Evaluation

Case managers highlighted evaluation and health education. First, they assessed and examined the content of the previous education and recorded and analyzed the patients' conditions, medication, diet, nutrition, rehabilitation exercises, and self-management. Second, they prepared the personalized health education plan, return visit plan, and rehabilitation plan, and used standardized courseware, educational videos, and health prescriptions to provide the patients with one-on-one health guidance. Finally, they sent the management tasks and educational contents to the phones of the patients for consolidating the learning in the hospital, as an outpatient, and at home.

Feedback

Patients can access their biochemical, physical, and chemical data as well as chronic disease–sensitive indicators in the hospital, as an outpatient, and at home for self-health management. Case managers can also perform online assessment, appraisal, and guidance via telephone, WeChat, and the chronic disease information system and record the data. Client mobile terminals can receive SMS text message alerts and the main interface of the chronic disease information system would display reminders of follow-up and return visits within ±7 days.

Settlement

If a patient was out of contact for 3 months, died, or refused to accept the treatment, case managers could settle the case.

Evaluation of the Effect of Implementing the Chronic Disease Management Information System

Evaluation Method

In accordance with case quality management indicators [20], two full-time case managers collected and evaluated data in the process of the follow-up procedure. To reduce the potential for evaluation bias, the case managers consistently communicated and learned to standardize the evaluation method. The cases were divided based on different chronic respiratory diseases case management models (ie, before and after use of the chronic disease information system). The following case management quality indicators were evaluated under the noninformation system management model (2019 - 2020) and under the chronic disease management information system model (2021 - 2022): number of managed cases, number of patients accepting routine follow-up services, follow-up visit rate, pulmonary function test rate, admission rate for acute exacerbations, chronic respiratory diseases knowledge awareness rate, and patient satisfaction. Excel sheets were used to acquire data prior to incorporation of the chronic disease management information system into the new information system.

Evaluation Indicators

The annual number of cases was calculated as the sum of the number of newly enrolled patients and the number of initially enrolled patients. The number of cases was calculated as the sum of the number of cases in different years. The number of routine follow-up visits represents the number of patients who completed the treatment plan. The follow-up visit rate was calculated as the number of completed follow-up visits in the year divided by the number of planned follow-up visits in the same year. The pulmonary function test rate was calculated as the number of pulmonary function tests completed for patients scheduled for follow-up during the year divided by the number of pulmonary function tests for patients scheduled for follow-up during the year. The admission rate for acute exacerbations was calculated as the number of recorded patients admitted to the hospital due to acute exacerbations divided by the total number of patients recorded. The chronic respiratory diseases knowledge awareness rate was determined by the number of people having sufficient knowledge divided by the total number of people tested. This knowledge indicator was based on the self-prepared chronic respiratory diseases knowledge test scale, which consists of 10 items determined using the Delphi method (following expert consultation) through review of the literature, including common symptoms, disease hazards, treatment medication, diet, living habits, exercise, negative habits affecting the disease, regular review items, effective methods for cough and sputum removal, appointments, and follow-ups. The content of the

questionnaire was refined by disease type, and the reviewers included 11 personnel with the title of Deputy Chief Nurse or above in the Internal Medicine Department of the hospital. The expert authority coefficients were 0.85 and 0.87 and the coordination coefficients were 0.50 and 0.67 for the two rounds of review, respectively; the χ^2 test showed a statistically significant value of P=.01. Patient satisfaction was assessed with a self-made questionnaire that showed good internal reliability (Cronbach α=0.78) and content validity (0.86). The questionnaire items included the reminder of return visits, practicability of health education content, and service attitude of medical staff; the full-time case managers surveyed the patients (or their caregivers) at the time of return visits after the third quarter of each year. Satisfaction items were rated using a 5-point Likert scale with a score of 1 - 5, and a mean ≥4 points for an individual indicated satisfaction. Patient satisfaction was then calculated as the number of satisfied patients divided by the total number of managed patients.

Statistical Analysis

SPSS 16.0 software was used for data analysis. The Mann-Whitney U test was performed to compare continuous variables between groups and the χ^2 test was performed to compare categorical variables between groups. P<.05 indicated that the difference was statistically significant.

Ethical Considerations

The study was conducted in accordance with the principles of the Declaration of Helsinki. This study received approval from the Ethics Committee of Yulin First People's Hospital (approval number: YLSY-IRB-RP-2024005). The study did not interfere with routine diagnosis and treatment, did not affect patients' medical rights, and did not pose any additional risks to patients. Therefore, after discussion with the Ethics Committee of Yulin First People's Hospital, it was decided to waive the requirement for informed consent from patients. Patients' personal privacy and data confidentiality have been upheld throughout the study.

Results

Characteristics of Patient Populations Before and After Implementation of the Information System

There was no significant difference in age and gender distributions in the patient populations that received care before and after implementation of the chronic disease management information system (Table 1).

Table . General characteristics of the patient populations under case management before and after use of the chronic disease management information system.

Characteristic		Before use (n=1024)	After use (n=2747)	χ^2 value	*df*	*P* value
Gender, n (%)				1.046	1	.31
	Men	677 (66.1)	1767 (64.3)			
	Women	347 (33.9)	980 (35.7)			
Age group (years), n (%)				0.997	3	.80
	<30	26 (2.6)	73 (2.7)			
	30-59	370 (36.1)	1013 (36.9)			
	60-79	510 (49.8)	118 (11.5)			
	>80	1322 (48.1)	339 (12.3)			

Comparison of Workload Before and After Implementation of the Information Management System

Before use of the system, 1808 cases were managed, with a mean of 603 (SD 137) cases having routine follow-up visits. After use of the system, 5868 cases were managed, with a mean of 2056 (SD 211) routine follow-up visits. Therefore, the number of managed cases and the number of follow-up visits significantly increased by 3.2 and 3.4 times, respectively, after use of the system (U=342.779; P<.001).

Comparison of Quality Indicators of Managed Cases Before and After Implementation of the Information System

The quality indicators in the two groups are summarized in Table 2. Compared with the corresponding indicators before use of the system, the follow-up visit rate increased by 50.2%, the pulmonary function test rate increased by 26.2%, the chronic respiratory diseases knowledge awareness rate increased by 20.1%, the retention rate increased by 16.3%, and the patient satisfaction increased by 9.6%; moreover, the admission rate for acute exacerbations decreased by 42.4%.

Table . Comparison of case management quality indicators before and after implementation of the chronic diseases information management system.

Quality indicators	Before use (n=1024), n (%)	After use (n=2747), n (%)	χ^2 value (*df*=1)	*P* value
Subsequent visit rate	209 (20.4)	1939 (70.6)	7.660	<.001
Lung function test achievement rate	190 (18.6)	1231 (44.8)	2.190	<.001
CRD[a] knowledge awareness rate	443 (43.3)	1742 (63.4)	1.243	<.001
Retention rate	787 (76.9)	2560 (93.2)	1.995	<.001
Acute exacerbation admission rate	663 (64.7)	613 (22.3)	5.999	<.001
Patient satisfaction	862 (84.2)	2577 (93.8)	86.190	.01

[a]CRD: chronic respiratory disease.

Discussion

Principal Findings

The main purpose of this study was to build a chronic disease management information system and apply it to the case management of chronic respiratory diseases. Our evaluation showed that the chronic disease management information system not only improves the efficiency and quality of case management but also has a benefit for maintaining the stability of the condition for patients with respiratory diseases, reduces the number of acute disease exacerbations, increases the rate of outpatient return, and improves patients' adherence with disease self-management. Thus, a chronic disease management information system is worth popularizing and applying widely.

Value of the HIS-Based Chronic Disease Management Information System

Chronic diseases constitute a significant public health issue in China. Public hospitals play important roles in the health service system, particularly large-scale public hospitals with the most advanced technologies, equipment, and enormous medical human resources, which can greatly aid in the diagnosis and treatment of diseases and also serve as important hubs for the graded treatment of chronic diseases. Moreover, a significant number of patients with chronic diseases visit large hospitals, making them important sources of big data on chronic diseases [21]. Adoption of an HIS-based chronic disease management information system can make full use of and exert the advantages of large-scale public hospitals in terms of labor, technology, and equipment in the diagnosis, treatment, and prevention of chronic diseases; enhance the cohesiveness of the case management team in chronic disease management; and achieve prehospital, in-hospital, and posthospital continuity of care for patients with chronic diseases. Overall, use of a chronic disease management information system can enhance the quality and efficiency of chronic disease management and lay a good foundation for teaching and research on chronic diseases.

Improved Efficiency of Case Management

China was relatively late in applying case management practices, and chronic disease management has traditionally been primarily conducted offline [14,20] or supplemented by management with apps and WeChat [7,8]. Traditional case management methods require case managers to manually search, record, store, query, count, and analyze information. This manual process necessitates substantial time and makes it challenging to realize a comprehensive, systematic, and dynamic understanding of patient information, resulting in a small number of managed cases and follow-up visits. With the application of information technology, use of an HIS-based chronic disease monitoring and case management system can automatically extract and integrate patient information, thereby increasing the efficiency of chronic disease management and reduce costs [4,22]. In this study, two case managers played leading roles both before and after implementation of the information system; however, compared with the situation before the use of the system, the numbers of both managed cases and of follow-up visits increased, reaching 3.2 and 3.4 times the preimplementation values, respectively. The information system can automatically obtain a patient's name and ID number based on the *ICD* code, which can expand the range of enrollment screening and appoint the register of patients as planned. In addition, the information system can automatically obtain outpatient, inpatient, and home medical information for the postillness life cycle management of patients. Moreover, the intuitive, clear, and dynamic indicator charts on the system can save a significant amount of time for diagnosis and treatment by medical staff, while the paperless office and online data-sharing functions can essentially solve the problem of managing files by case managers to ultimately enhance efficiency.

Improved Quality of Case Management

According to evidence-based medicine, the seven steps of case management represent the optimal clinical pathway [10-14,22]. In this study, the concept of an Internet-Plus medical service was introduced; that is, the chronic disease management information system was established based on the HIS data and case management model [22] and the function of a mobile medical terminal app was incorporated in the system [6,7]. Compared with the noninformation system case management model, this system has several advantages. First, owing to the swift management mode, it can overcome the limitations of time and space [4-8]. Second, multichannel health education and communication can enhance patients' knowledge and skills, as well as their compliance with self-management, based on diversified forms of image data such as graphics and audio [6,22]. Third, the use of intelligent management can remind

doctors and patients to complete management work and follow-up visits as planned, and to perform intelligent pushes of patient outcome indicators to improve confidence in the treatment [22]. Fourth, this system enables information sharing and big data analysis, as well as multidisciplinary diagnosis and treatment based on the matching of doctor-patient responsibility management, which can be more conducive to the precise health management of patients.

Compared with the traditional case management model, information-based case management significantly increased the follow-up visit rate, lung function test rate, chronic respiratory diseases knowledge awareness rate of patients, patient satisfaction rate, and retention rate. Among these indicators, the follow-up visit rate and lung function test rate represent aspects related to the patients' own management of their condition [1]. The results of this study are consistent with previous findings related to information-based management of chronic diseases in China, demonstrating that such a management system was more conducive to planned, systematic, and personalized education and follow-up by the case management team, thereby promoting the virtuous cycle of compliance with self-management and reducing the number of acute exacerbations among patients with chronic respiratory diseases, ultimately enhancing the precision of medical resource allocation and hospital management [22,23].

Helping Patients With COPD Maintain Stability of Their Condition

The admission rate for acute exacerbations serves as a common indicator of the quality of the treatment of chronic respiratory diseases [23]. The deployment of a clinical pathway–based hospital case management information system significantly reduced the admission rate for acute exacerbations and enhanced the quality of treatment for chronic respiratory diseases, indicating its high clinical significance. There are several reasons for these observed benefits. First, home care and self-management are essential in the management of chronic respiratory diseases. The information-based case management model improved the patients' knowledge and skills along with their compliance with self-management. Consequently, the standardized self-management process helped to reduce the number of acute exacerbations of chronic respiratory diseases and thus lowered the admission rate. Second, the information-based case management model increased the regular return rate, which allowed the medical staff to identify the potential risk factors for acute exacerbations in a timely manner, deal with them when they occur, and prepare personalized treatment plans and precise health management schemes. This consequently enabled adjustment of treatment schemes in real time, reduced the number of admissions due to acute exacerbations, and lowered the readmission rate. For hospitals interested in implementing a similar model, we suggest first conducting a detailed review of the current situation prior to making adequate changes based on the relevant disease and patient populations.

Consequently, the HIS-based case information management model could improve efficiency, enhance the quality of case management, and aid in stabilizing the conditions of patients with chronic respiratory diseases. In contrast to the hospital case management information system reported by Yuan et al [22], the system described in this study includes a personal terminal app. Previous studies confirmed that a stand-alone mobile health app could improve patient compliance and disease control [6-8]; thus, whether this system can be used to manage specialized disease cohorts for patients with chronic diseases remains to be determined. In this study, the effect on the retention rate of patients was confirmed; however, the overall operational indicators for the diagnosis and treatment of chronic diseases should be further determined.

Conclusion

With the advancement of information technology, the internet and medical technology have been applied to the management of chronic diseases. As an information-based platform for the case management of patients with chronic respiratory diseases, a newly developed chronic disease management information system was introduced in this study. This system is capable of designing the follow-up time registration, follow-up content, approaches, methods, quality control, and feedback process for a single chronic respiratory disease via the single-disease clinical pathway following the case management process (enrollment, assessment, planning, implementation, feedback, and evaluation). Use of this system can encourage patients with chronic respiratory diseases to adhere to regular follow-up and form an outpatient-inpatient-home chronic disease management strategy. This can help in reducing the admission rate for acute exacerbations, increase the return visit rate, and improve the correctness and compliance of home self-management of patients with chronic respiratory diseases. Owing to these benefits, wide adoption of such information systems for the management of chronic diseases can offer substantial economic and social value.

Acknowledgments

We are particularly grateful to all the people who provided help with our article. This study was supported by a grant from Yulin City Science and Technology Planning Project (20202002).

Data Availability

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

Authors' Contributions

YML, YZX, XLS, and YXL designed this study. XLS and XJC wrote the draft of the paper. YML, YZX, and YC contributed final revisions to the article. XJC, HC, and YC collected the data. XJC, YML, YXL, and HC performed the statistical analysis. YML received funding support. All authors read and approved the final draft of the article.

Conflicts of Interest

None declared.

References

1. Anaev EK. Eosinophilic chronic obstructive pulmonary disease: a review. Ter Arkh 2023 Oct 11;95(8):696-700. [doi: 10.26442/00403660.2023.08.202316] [Medline: 38158908]
2. Shakeel I, Ashraf A, Afzal M, et al. The molecular blueprint for chronic obstructive pulmonary disease (COPD): a new paradigm for diagnosis and therapeutics. Oxid Med Cell Longev 2023 Dec;2023:2297559. [doi: 10.1155/2023/2297559] [Medline: 38155869]
3. Morimoto Y, Takahashi T, Sawa R, et al. Web portals for patients with chronic diseases: scoping review of the functional features and theoretical frameworks of telerehabilitation platforms. J Med Internet Res 2022 Jan 27;24(1):e27759. [doi: 10.2196/27759] [Medline: 35084355]
4. Donner CF, ZuWallack R, Nici L. The role of telemedicine in extending and enhancing medical management of the patient with chronic obstructive pulmonary disease. Medicina 2021 Jul 18;57(7):726. [doi: 10.3390/medicina57070726] [Medline: 34357007]
5. Wu F, Burt J, Chowdhury T, et al. Specialty COPD care during COVID-19: patient and clinician perspectives on remote delivery. BMJ Open Respir Res 2021 Jan;8(1):e000817. [doi: 10.1136/bmjresp-2020-000817] [Medline: 33414261]
6. Hallensleben C, van Luenen S, Rolink E, Ossebaard HC, Chavannes NH. eHealth for people with COPD in the Netherlands: a scoping review. Int J Chron Obstruct Pulmon Dis 2019 Jul;14:1681-1690. [doi: 10.2147/COPD.S207187] [Medline: 31440044]
7. Gokalp H, de Folter J, Verma V, Fursse J, Jones R, Clarke M. Integrated telehealth and telecare for monitoring frail elderly with chronic disease. Telemed J E Health 2018 Dec;24(12):940-957. [doi: 10.1089/tmj.2017.0322] [Medline: 30129884]
8. McCabe C, McCann M, Brady AM. Computer and mobile technology interventions for self-management in chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2017 May 23;5(5):CD011425. [doi: 10.1002/14651858.CD011425.pub2] [Medline: 28535331]
9. Briggs AM, Persaud JG, Deverell ML, et al. Integrated prevention and management of non-communicable diseases, including musculoskeletal health: a systematic policy analysis among OECD countries. BMJ Glob Health 2019;4(5):e001806. [doi: 10.1136/bmjgh-2019-001806] [Medline: 31565419]
10. Franek J. Home telehealth for patients with chronic obstructive pulmonary disease (COPD): an evidence-based analysis. Ont Health Technol Assess Ser 2012;12(11):1-58. [Medline: 23074421]
11. Shah A, Hussain-Shamsy N, Strudwick G, Sockalingam S, Nolan RP, Seto E. Digital health interventions for depression and anxiety among people with chronic conditions: scoping review. J Med Internet Res 2022 Sep 26;24(9):e38030. [doi: 10.2196/38030] [Medline: 36155409]
12. Sugiharto F, Haroen H, Alya FP, et al. Health educational methods for improving self-efficacy among patients with coronary heart disease: a scoping review. J Multidiscip Healthc 2024 Feb;17:779-792. [doi: 10.2147/JMDH.S455431] [Medline: 38410523]
13. Metzendorf MI, Wieland LS, Richter B. Mobile health (m-health) smartphone interventions for adolescents and adults with overweight or obesity. Cochrane Database Syst Rev 2024 Feb 20;2(2):CD013591. [doi: 10.1002/14651858.CD013591.pub2] [Medline: 38375882]
14. Reig-Garcia G, Suñer-Soler R, Mantas-Jiménez S, et al. Assessing nurses' satisfaction with continuity of care and the case management model as an indicator of quality of care in Spain. Int J Environ Res Public Health 2021 Jun 19;18(12):6609. [doi: 10.3390/ijerph18126609] [Medline: 34205373]
15. Aggelidis X, Kritikou M, Makris M, et al. Tele-monitoring applications in respiratory allergy. J Clin Med 2024 Feb 4;13(3):898. [doi: 10.3390/jcm13030898] [Medline: 38337592]
16. Seid A, Fufa DD, Bitew ZW. The use of internet-based smartphone apps consistently improved consumers' healthy eating behaviors: a systematic review of randomized controlled trials. Front Digit Health 2024;6:1282570. [doi: 10.3389/fdgth.2024.1282570] [Medline: 38283582]
17. Verma L, Turk T, Dennett L, Dytoc M. Teledermatology in atopic dermatitis: a systematic review. J Cutan Med Surg 2024;28(2):153-157. [doi: 10.1177/12034754231223694] [Medline: 38205736]
18. Tański W, Stapkiewicz A, Szalonka A, Głuszczyk-Ferenc B, Tomasiewicz B, Jankowska-Polańska B. The framework of the pilot project for testing a telemedicine model in the field of chronic diseases - health challenges and justification of the project implementation. Pol Merkur Lekarski 2023;51(6):674-681. [doi: 10.36740/Merkur202306115] [Medline: 38207071]

19. Popp Z, Low S, Igwe A, et al. Shifting from active to passive monitoring of Alzheimer disease: the state of the research. J Am Heart Assoc 2024 Jan 16;13(2):e031247. [doi: 10.1161/JAHA.123.031247] [Medline: 38226518]
20. Sagare N, Bankar NJ, Shahu S, Bandre GR. Transforming healthcare: the revolutionary benefits of cashless healthcare services. Cureus 2023 Dec;15(12):e50971. [doi: 10.7759/cureus.50971] [Medline: 38259368]
21. Noncommunicable Diseases, Rehabilitation and Disability (NCD), Surveillance, Monitoring and Reporting (SMR) WHO Team. Noncommunicable diseases progress monitor. : World Health Organization; 2017 URL: https://www.who.int/publications/i/item/9789241513029 [accessed 2024-05-09]
22. Yuan W, Zhu T, Wang Y, et al. Research on development and application of case management information system in general hospital. Nurs Res 2022;36(12):2251-2253.
23. 2020 GOLD report. Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease. : Global Initiative for Chronic Obstructive Lung Disease; 2020 URL: https://goldcopd.org/gold-reports/ [accessed 2024-05-09]

Abbreviations

COPD: chronic obstructive pulmonary disease
HIS: hospital information system
***ICD*:** *International Classification of Diseases*

Edited by C Lovis; submitted 15.06.23; peer-reviewed by KM Kuo; revised version received 14.04.24; accepted 17.04.24; published 19.06.24.

Please cite as:
Xiao YZ, Chen XJ, Sun XL, Chen H, Luo YX, Chen Y, Liang YM
Effect of Implementing an Informatization Case Management Model on the Management of Chronic Respiratory Diseases in a General Hospital: Retrospective Controlled Study
JMIR Med Inform 2024;12:e49978
URL: https://medinform.jmir.org/2024/1/e49978
doi:10.2196/49978

Original Paper

Value of Electronic Health Records Measured Using Financial and Clinical Outcomes: Quantitative Study

Shikha Modi[1,2], MBA, PhD; Sue S Feldman[2], RN, MEd, PhD; Eta S Berner[2], EdD; Benjamin Schooley[3], PhD; Allen Johnston[4], PhD

[1]The University of Alabama in Huntsville, Huntsville, AL, United States

[2]The University of Alabama at Birmingham, Birmingham, AL, United States

[3]Brigham Young University, Provo, UT, United States

[4]Department of Information Systems, Statistics, and Management Science, The University of Alabama, Tuscaloosa, AL, United States

Corresponding Author:
Shikha Modi, MBA, PhD
The University of Alabama in Huntsville
1610 Ben Graves Dr NW
Huntsville, AL, 35816
United States
Phone: 1 2568242437
Email: ssm0031@uah.edu

Abstract

Background: The Health Information Technology for Economic and Clinical Health Act of 2009 was legislated to reduce health care costs, improve quality, and increase patient safety. Providers and organizations were incentivized to exhibit meaningful use of certified electronic health record (EHR) systems in order to achieve this objective. EHR adoption is an expensive investment, given the resources and capital that are invested. Due to the cost of the investment, a return on the EHR adoption investment is expected.

Objective: This study performed a value analysis of EHRs. The objective of this study was to investigate the relationship between EHR adoption levels and financial and clinical outcomes by combining both financial and clinical outcomes into one conceptual model.

Methods: We examined the multivariate relationships between different levels of EHR adoption and financial and clinical outcomes, along with the time variant control variables, using moderation analysis with a longitudinal fixed effects model. Since it is unknown as to when hospitals begin experiencing improvements in financial outcomes, additional analysis was conducted using a 1- or 2-year lag for profit margin ratios.

Results: A total of 5768 hospital-year observations were analyzed over the course of 4 years. According to the results of the moderation analysis, as the readmission rate increases by 1 unit, the effect of a 1-unit increase in EHR adoption level on the operating margin decreases by 5.38%. Hospitals with higher readmission payment adjustment factors have lower penalties.

Conclusions: This study fills the gap in the literature by evaluating individual relationships between EHR adoption levels and financial and clinical outcomes, in addition to evaluating the relationship between EHR adoption level and financial outcomes, with clinical outcomes as moderators. This study provided statistically significant evidence ($P<.05$), indicating that there is a relationship between EHR adoption level and operating margins when this relationship is moderated by readmission rates, meaning hospitals that have adopted EHRs could see a reduction in their readmission rates and an increase in operating margins. This finding could further be supported by evaluating more recent data to analyze whether hospitals increasing their level of EHR adoption would decrease readmission rates, resulting in an increase in operating margins. Hospitals would incur lower penalties as a result of improved readmission rates, which would contribute toward improved operating margins.

(JMIR Med Inform 2024;12:e52524) doi:10.2196/52524

KEYWORDS

acceptance; admission; adoption; clinical outcome; cost; economic; EHR adoption; EHR; electronic health record; finance; financial outcome; financial; health outcome; health record; hospital; hospitalization; length of stay; margin; moderation analysis;

multivariate; operating margin; operating; operation; operational; profit; project management; readmission rate; readmission; total margin; value analysis; value engineering; value management

Introduction

Overview

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 was legislated to reduce health care costs, improve quality, and increase patient safety [1]. Providers and organizations were incentivized to exhibit meaningful use of certified electronic health record (EHR) systems in order to achieve this objective [1]. The HITECH Act was based on the "triple aim" of health care, which consisted of reducing costs, improving the experience of care, and improving population health, and the HITECH Act contributed to the importance of EHRs [2]. Physicians and hospitals that adopted and used certified EHRs as described in federally defined "meaningful use" criteria were awarded approximately US $27 billion in incentives [3] for eligible providers.

EHR adoption is an expensive investment, given the resources and capital that are invested [4,5]. Due to the cost of the investment, a return on the EHR adoption investment is expected. Usually, a return on adoption is measured by calculating net profit and dividing the net profit by net investment [6]. Calculating a return on investment for EHR adoption requires considering the size of the organization, the extent of the EHR adoption, and profit or improvement in terms of both the financial and clinical outcomes perspectives. Given the complex process of calculating return on investment for EHR adoption, this study evaluates return on investment in terms of how it yields value to the adopting entity. Value from the health care perspective has been described in terms of dollars (financial), productivity (clinical), effectiveness (clinical) [7], cost savings (financial) [8], improvement in clinical decisions (clinical; Rudin et al [9]), supporting triage decisions (clinical; Rudin et al [9]), supporting collaborations among the providers (clinical; Rudin et al [9]), increased productivity (financial and clinical) [9], etc. However, a gap exists in that the return on investment is not analyzed in terms of financial and clinical outcomes combined. Additionally, current literature does not review EHR adoption in terms of level of EHR adoption but rather as a binary variable of "adopted" or "not adopted." This study addresses these gaps by including a combination of both financial and clinical outcomes in a conceptual model and reviewing EHR adoption in terms of levels of EHR adoption.

The value of health IT, of which EHRs are a subset, can depend on the stakeholder and context [10-12]. Looking at value from the stakeholder perspective, for the hospital, EHR value may be reviewed in terms of improved revenue and reduced cost (outcomes); for patients, value may be to improve health and prevent illness (outcomes); for providers, value may be to reduce errors and provide efficient care (process and outcomes); and for government, it may be to improve population health through timely public health reporting and population well-being (process and outcomes) [7]. Hence, given the frequent use of different outcome categories in the literature used to measure value, this study focuses on outcomes as the main value construct and investigates value in terms of different tangible outcomes, such as financial and clinical outcomes. This study examined how EHR adoption levels are associated with value in terms of financial and clinical outcomes combined in 1 model. To address this question, this study investigated the relationship between EHR adoption levels and financial and clinical outcomes by combining both financial and clinical outcomes into 1 conceptual model.

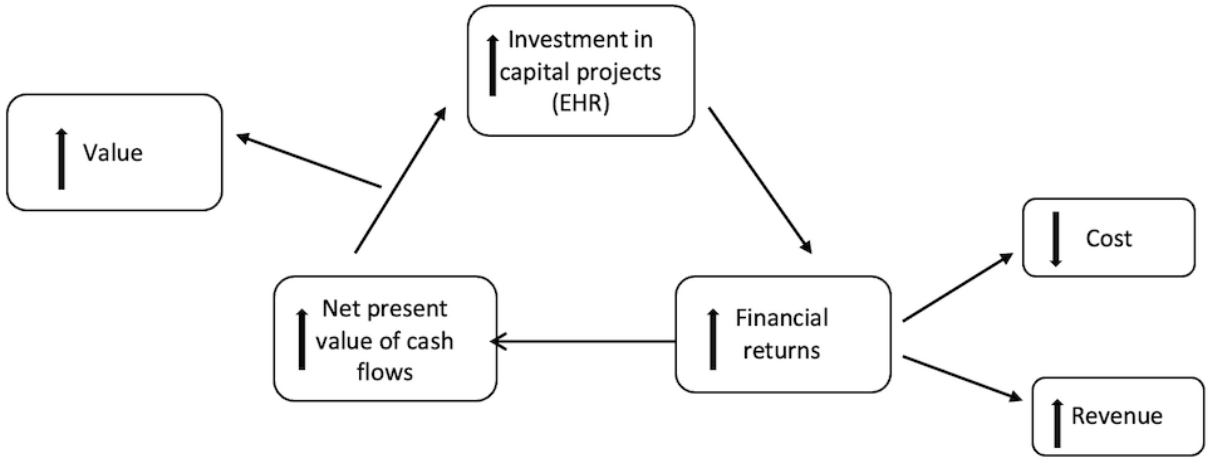
Conceptual Framework and Hypotheses

This study used the corporate financial theory of the firm [13] to guide the evaluation of the relationship between EHR adoption and financial and clinical outcomes. The corporate financial theory of the firm (Figure 1) indicates that the value of the firm, or in this case, the health care entity, is expected to be in alignment with the discounted cash flows from the investments, such as EHRs [13]. This theory indicates that a capital investment, such as EHR adoption, increases the value of the firm as it contributes toward an increase in the net present value of cash flows [13]. Multiple studies have supported the notion that EHR investments improve the value of a hospital through improved financial outcomes by way of a reduction in cost or improved revenues [4,14,15].

A study conducted by Collum et al [4] used this theory to determine an association between EHR adoption and financial outcomes (measured as profit margins and return on assets). The findings from this study indicated that financial returns depend on how long it takes for a hospital's EHR system to achieve full functionality [4], meaning it is important to consider the time variable when reviewing the outcomes of EHR adoption.

Additionally, there have been several studies that have analyzed the relationship between EHR adoption and financial outcomes without using the corporate financial theory of the firm as their guiding framework. Some of the studies from the current literature exhibited a trend that EHR adoption and financial outcomes have a nonlinear relationship [16,17], and some of the studies indicated that EHR adoption resulted in improved financial outcomes for health care organizations that adopted it over time [14,18].

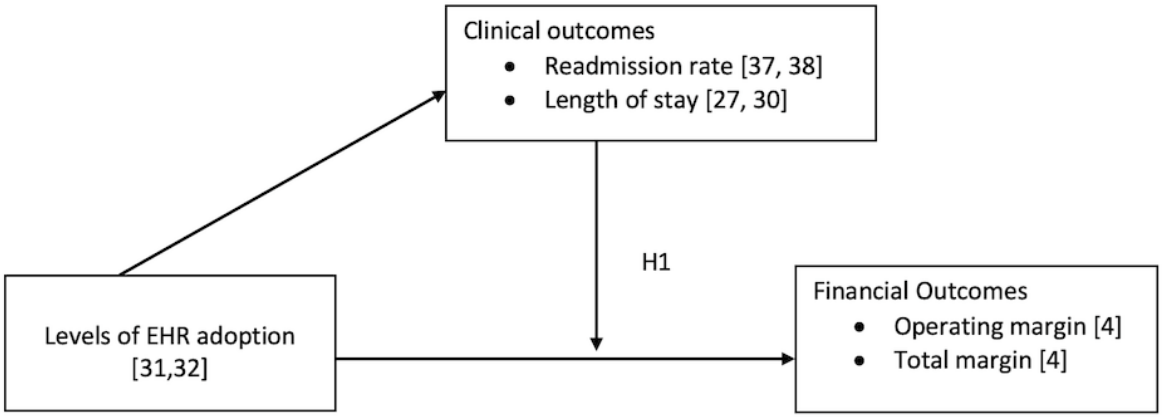
Figure 1. The corporate financial theory of the firm. EHR: electronic health record.



The literature suggests that improvement in costs and revenues is the result of improved clinical outcomes such as reduction of redundant tests [19], reduction of medication and hospital bed-related costs [20], improved ability to capture charges [15], and improved decision support systems [21]. Since this study focuses on combining both financial and clinical outcomes into 1 conceptual model, for the purpose of this study, the capital project investment (EHR adoption in this case) and improvement in financial returns (financial outcomes), tenets of the corporate financial theory of the firm, with an addition of the clinical outcomes, are integrated into a conceptual framework.

The purpose of this conceptual framework (Figure 2 [4,22-27]) is to determine if the previously stated overarching research question of "How is electronic health record adoption associated with value in terms of financial and clinical outcomes?" is supported by the following hypothesis: "The relationship between levels of EHR adoption and financial outcomes (both operating margin and total margin) is moderated by clinical outcomes (readmission rate and length of stay [LOS]) that are also affected by levels of EHR adoption (Figure 2).

Figure 2. Electronic health record (EHR) value analysis conceptual framework.



Methods

Data for this study were retrieved from multiple sources, including the Health Care Provider Cost Reporting Information System, the American Hospital Association (AHA) Annual Survey, the AHA IT Supplement Survey, and Health Care Analytics from Leavitt Partners. The study used a longitudinal design from 2014 to 2017 with 5897 hospital-year observations. Measures were divided into 2 groups: financial and clinical. Financial outcome variables were measured or operationalized using 2 variables (operating margin and total margin) that have been used in the health care literature to measure the profitability of hospitals after EHR adoption. The variables describing clinical outcomes are LOS and readmission rates, as these variables have an impact on the financial performance of the hospital [28,29]. The variables describing the financial outcomes are operating margin and total margin, as these measures include both costs and revenues described in the corporate financial theory of the firm [4,13]. The dependent variables used in this study (operating margins, total margins, LOS, and readmission rates) are not comprehensive in terms of measuring financial

and clinical outcomes for a hospital; however, for the purpose of this study, these variables are considered sufficient, given their potential association with one another.

Dependent Variables

Financial Outcome Variables

In order to gain an understanding of the financial performance of acute care hospitals, profitability ratios are the most frequently used measures [30]; hence, this study included operating margin and total margin as variables representing financial outcomes. Operating margin captures the expenses and revenues related to hospital operations. Total margin measures or captures operating and nonoperating expenses and revenues. The operating margin was calculated by dividing net patient revenues less total operating expenses by net patient revenues and multiplying the ratio by 100. The total margin variable is calculated by dividing net income by total patient revenue. The financial outcome variables are retrieved from the AHA Annual Survey (2014-2017).

Clinical Outcome Variables

Clinical outcomes were measured using LOS and readmission rates. Daniel et al [22] and Schreiber and Shaha [31] reported an intersection of financial and clinical outcomes as a result of EHR adoption and focused on LOS. These studies reported an improvement in LOS due to EHR adoption, resulting in lower plan premiums for patients [22] and costs [31]. Readmission rates are a part of the value-based purchasing program, and depending on the readmission rate, hospitals are penalized on a yearly basis, hence impacting hospital costs [28]. The readmission rates were measured for 6 conditions or procedures, as patients with these conditions are more likely to be readmitted to the hospital. These conditions are: acute myocardial infarction, chronic obstructive pulmonary disease, heart failure, pneumonia, coronary artery bypass graft surgery, and elective primary total hip arthroplasty and total knee arthroplasty [32]. LOS captures the number of days a patient spent in the hospital. Readmission rates indicate whether patients are readmitted to the hospital within 30 days of being discharged. The average LOS and readmission rates can be considered to be indicators of clinical quality outcomes by way of clinical quality measures [28]. Ben-Assuli et al [33] and Lee et al [34] have indicated improvements in average LOS and readmission rates as results of EHR adoption. To confirm these findings for the most recent data, this study analyzes how EHR adoption influences both average (LOS) and readmission rates for the selected sample.

The LOS variable is measured as the average number of days a patient stays in one hospital. The readmission rate variable is measured as the readmission rate payment adjustment factor. The full-year payment adjustment factor is based on data from the fiscal year Hospital Readmissions Reduction Program performance period (ie, July 1, 2014, to June 30, 2017). The minimum payment adjustment factor is 0.97 (ie, 3% maximum penalty). The maximum payment adjustment factor is 1 (ie, no penalty). Hospitals with higher payment adjustment factors have lower penalties [32].

Independent Variables

The level of EHR adoption is considered the major explanatory variable in this study. Hospitals are required to report the extent of adoption of each of the 28 EHR functions to the AHA IT Supplement Survey. The 28 EHR functions can be characterized into 5 different categories: clinical documentation, results viewing, computerized order entry, decision support, and bar coding. Hospitals indicate if each function is implemented in all units, 1 unit, or is in some stage of planning. A study conducted by Everson et al [23] emphasizes the reliability and validity of measuring hospital adoption of EHR with these 28 items.

In order to look at the extent of EHR adoption, Adler-Milstein et al [24] created a continuous EHR adoption measure for each hospital in each year in which they responded to the AHA IT Supplement Survey. The continuous measure was constructed as follows: for each function that was implemented in all units, a hospital received 2 points, and for each function that was implemented in at least 1 unit, a hospital received 1 point. According to the calculations, the total possible EHR adoption score ranged from 0 to 56. In order to improve interpretability, the measure was scaled by dividing each hospital's total score by 56, which yielded an EHR score ranging from 0 to 1. This strategy will be replicated in this study and applied to the EHR adoption level [24].

Control Variables

Control variables for this study include time-variant variables such as competition and payer mix. Control variables are identified based on elements that may influence the level of EHR adoption or hospital financial and clinical outcomes [4]. Since this study uses panel data, which accounts for changes in financial outcomes within hospitals due to changes in levels of EHR adoption, it is not essential to control for time-invariant hospital characteristics such as size of the hospital, ownership, system affiliation, and teaching status. For the purpose of this study, time-variant components that may change over the years, such as competition and payer mix, are considered control variables [4].

The competition construct was operationalized using the Herfindahl-Hirschman Index (HHI), which measures the concentration of an industry in a designated market. HHI was measured in terms of discharges for the health service area. Payer mix was measured using the proportion of inpatient days that were related to Medicare and Medicaid patients (Medicare percentage = total facility Medicare days/total inpatient days, and Medicaid percentage = total facility Medicaid days/total inpatient days). The AHA Annual Survey was used to collect the HHI and payer mix data.

Analysis

The unit of analysis for this study is at the hospital level. To demonstrate the appropriateness of the variables, univariate statistics and bivariate analyses were conducted. Bivariate statistics were generated for both independent and dependent variables of interest. Pairwise correlation analysis was conducted at the significance level of $P<.05$ in order to examine pairwise correlation coefficients between the continuous variables.

Multivariate relationships between different levels of EHR adoption and financial and clinical outcomes, along with the time-variant control variables, were examined using moderation analysis with a longitudinal fixed effects model [35]. Since it is unknown as to when hospitals begin experiencing improvements in financial outcomes, additional analysis was conducted using a 1- or 2-year lag for profit margin ratios [4]. Statistical significance was noted at the significance levels of $P<.10$, $P<.05$, and $P<.01$, and all statistical analyses were conducted in Stata (version 16; StataCorp).

Longitudinal Fixed Effects Moderation Analysis Model

A longitudinal fixed effects model with moderation analysis was used to analyze the multivariate relationships between different levels of EHR adoption and financial and clinical outcomes, along with the time variant control variables.

$$y_{it} = \beta_1 X_{it1} + \beta_2 X_{it2} + \beta_3 X_{it1} X_{it2} + Z_{it}\lambda + \alpha_i + \mu_{it}$$

In this equation, y_{it} is the dependent variable (financial or clinical outcomes), i = hospital, and t = time. β_1 is the coefficient for the main independent variable (levels of EHR adoption), X_{it1}. β_2 is the coefficient for the moderator variable (clinical outcomes), X_{it2}. β_3 is the coefficient for the interaction of the independent variable (levels of EHR adoption) and moderator variable (clinical outcomes), $X_{it1}X_{it2}$. $Z_{it}\lambda$ represents all control variables (competition, payer mix, and years of observation). α_i is the unknown intercept for a vector of hospitals. And μ_{it} is the error term.

The hypothesis, that the relationship between EHR adoption and financial outcomes is moderated by clinical outcomes, was tested using multiple models. The models and their use are outlined in Textbox 1.

Textbox 1. Analytic models and their use.

Model 1

Determine the association between levels of electronic health record (EHR) adoption and operating margin moderated by length of stay (LOS) with the operating margins from the same year.

Model 2

Determine the association between levels of EHR adoption and operating margin moderated by readmission rates with the operating margins from the same year.

Model 3

Determine the association between levels of EHR adoption and total margin moderated by LOS with the total margins from the same year.

Model 4

Determine the association between levels of EHR adoption and total margin moderated by readmission rates with the total margins from the same year.

Model 5

Determine the association between levels of EHR adoption and operating margin moderated by LOS with a 1-year lag in the operating margins.

Model 6

Determine the association between levels of EHR adoption and operating margin moderated by LOS with a 2-year lag in the operating margins.

Model 7

Determine the association between levels of EHR adoption and operating margin moderated by readmission rates with a 1-year lag in the operating margins.

Model 8

Determine the association between levels of EHR adoption and operating margin moderated by readmission rates with a 2-year lag in the operating margins.

Model 9

Determine the association between levels of EHR adoption and total margin moderated by LOS with a 1-year lag in the total margins.

Model 10

Determine the association between levels of EHR adoption and total margin moderated by LOS with a 2-year lag in the total margins.

Model 11

Determine the association between levels of EHR adoption and total margin moderated by readmission rates with a 1-year lag in the total margins.

Model 12

Determine the association between levels of EHR adoption and total margin moderated by readmission rates with a 2-year lag in the total margins.

Ethical Considerations

This study was approved by the University of Alabama at Birmingham institutional review board (300003241).

Results

Overview

Descriptive statistics of acute care hospitals for the years 2014-2017 are displayed in Table 1. For acute care hospitals, average EHR adoption levels showed little variability across each observed year (approximately 0.89 for each observed year). Hospitals observed a steady decrease in average operating margin from 2014 (0.07%) to 2017 (0.057%). The average total margin across hospitals showed a decrease for 2015 (0.005%) compared with 2014 (1.014%), followed by a steady increase across years 2016 (1.136%) and 2017 (0.951%). An increase in LOS was observed for the years 2016 and 2017 (approximately 7.9 days for the year 2017 vs 3.9 days for the year 2014). Average readmission rates remained somewhat steady across all 4 observation years (approximately 0.99 for each observed year).

Table 1. Descriptive statistics of variables (N=5678 hospital-year observations).

Variables	2014 (n=1420)	2015 (n=1453)	2016 (n=1393)	2017 (n=1412)
Levels of EHR[a] adoption, mean (SD)	0.871 (0.127)	0.890 (0.121)	0.899 (0.127)	0.917 (0.102)
Operating margin, mean (SD)	0.070 (0.122)	0.065 (0.132)	0.06 (0.140)	0.057 (0.136)
Total margin, mean (SD)	1.014 (2.847)	0.005 (26.4)	1.136 (7.217)	0.951 (1.129)
Average length of stay (days), mean (SD)	3.911 (1.134)	3.881 (0.954)	7.87 (153.3)	7.945 (160.4)
Readmission rate, mean (SD)	0.998 (0.003)	0.995 (0.006)	0.995 (0.006)	0.994 (0.007)
Market competition (HHI[b]) in terms of discharges, mean (SD)	0.101 (0.199)	0.086 (0.157)	0.088 (0.172)	0.098 (0.193)
Medicare percentage, mean (SD)	0.512 (0.140)	0.518 (0.128)	0.518 (0.130)	0.521 (0.124)
Medicaid percentage, mean (SD)	0.197 (0.120)	0.202 (0.115)	0.203 (0.114)	0.204 (0.112)
Beds (n), mean (SD)	257 (231)	256 (229)	254 (232)	255 (236)
Ownership, n (%)				
Nongovernment not-for-profit	1105 (77.76)	1145 (78.8)	1177 (78.31)	1198 (78.82)
Investor-owned for-profit	294 (20.69)	295 (20.30)	311 (20.69)	305 (20.07)
Government nonfederal	22 (1.55)	13 (0.89)	15 (1)	17 (1.12)
Affiliation, n (%)				
Yes	584 (47.29)	660 (51.36)	687 (51.58)	731 (56.67)
No	651 (52.71)	625 (48.64)	645 (48.42)	559 (43.33)
Teaching status, n (%)				
Yes	560 (39.41)	569 (39.16)	595 (39.59)	599 (39.41)
No	861 (60.59)	884 (60.84)	908 (60.41)	921 (60.59)

[a]EHR: electronic health record.

[b]HHI: Herfindahl-Hirschman Index.

For time-variant control variables, the average HHI in terms of discharges across all 4 years was approximately 0.093. HHI values range from 0 to 1, where an HHI value closer to 1 means monopolistic markets, or more market share, and an HHI value closer to 0 means highly competitive markets, or less market share. For the sample used in this study, the markets appear to be highly competitive. In terms of payer mix, the Medicare percentage was similar across all 4 years (average of 0.52). Similarly, the Medicaid percentage was also similar across all 4 years (average of 0.20).

For organizational characteristics, bed size was somewhat similar across all hospitals for all observed years (approximately 255 beds per hospital). In terms of ownership status of the sample hospitals, a majority of the hospitals were nongovernment, not-for-profit hospitals (1105/1421, 78%), followed by investor-owned for-profit hospitals (294/1421, 20%) and government nonfederal hospitals (22/1421, 1.5%). In terms of system affiliation, approximately half the hospitals were affiliated with a system, and the other half were not. For teaching status, a majority of the hospitals did not hold a teaching status (861/1421, 61%).

According to the bivariate statistical analysis (Table 2), at the significance level of $P<.05$, levels of EHR adoption exhibit a positive correlation with operating margin at a magnitude of 0.0978. At the significance level of $P<.05$, readmission rate and levels of EHR adoption are negatively correlated at the magnitude of 0.0321. Even though the magnitudes are close to

0, these relationships are statistically significant at the significance level of $P<.05$.

Table 2. Bivariate analysis of variables.

Dependent variables	Independent variables: levels of EHR[a] adoption (correlation coefficients)
Operating margin	0.0978[b]
Total margin	–0.0142
Average length of stay	0.0039
Readmission rate	–0.0321[b]

[a]EHR: electronic health record.

[b]$P<.05$.

This study tested the following hypothesis that was derived from the EHR value analysis conceptual framework (Figure 2): "The relationship between EHR adoption and financial outcomes is moderated by clinical outcomes." Tables 3 and 4 provide details relative to the hypothesis.

Table 3. Fixed effects with regression analysis.

Variables	Model 1	Model 2	Model 3	Model 4
	OM[a]–LOS[b]–levels of EHR[c] adoption (Prob>F=0.0828)	OM–RR[d]–levels of EHR adoption (Prob>F=0.0116)	TM[e]–LOS–levels of EHR adoption (Prob>F=0.4532)	TM–RR–levels of EHR adoption (Prob>F=0.3388)
Levels of EHR adoption	–0.020	5.335[f]	–4.961	415.2
Dependent variables				
Average LOS	0.000	N/A[g]	–0.002	N/A
RR	N/A	4.375[h]	N/A	431.6
Levels of EHR adoption and average LOS	-0.000	N/A	0.001	N/A
Levels of EHR adoption and RR	N/A	–5.384[f]	N/A	–422.3
Control variables				
Market competition (HHI[i])	0.082	0.078	3.148	2.959
Medicare percentage	–0.009	–0.013	0.699	0.937
Medicaid percentage	–0.026	–0.026	1.211	1.343
Years				
2014	Reference	Reference	Reference	Reference
2015	–0.005	–0.007[f]	–1.001	–0.848
2016	–0.008	–0.009[f]	0.388	0.569
2017	–0.008	–0.011[j]	0.243	0.460

[a]OM: operating margin.

[b]LOS: length of stay.

[c]EHR: electronic health record.

[d]RR: readmission rate.

[e]TM: total margin.

[f]$P<.05$.

[g]N/A: not applicable.

[h]$P<.10$.

[i]HHI: Herfindahl-Hirschman Index.

[j]$P<.001$.

Table 4. Regression analysis with fixed effects for lagged variables.

Variables	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
	OM[a]–LOS[b]–levels of EHR[c] adoption with 1-year lag (Prob>F=0.0047)	OM–LOS–levels of EHR adoption with 2-year lag (Prob>F=0.0271)	OM–RR[d]–levels of EHR adoption with 1-year lag (Prob>F=0.0010)	OM–RR–levels of EHR adoption with 2-year lag (Prob>F=0.0202)	TM[e]–LOS–levels of EHR adoption with 1-year lag (Prob>F=0.6885)	TM–LOS–levels of EHR adoption with 2-year lag (Prob>F=0.6738)	TM–RR–levels of EHR adoption with 1-year lag (Prob>F=0.6492)	TM–RR– levels of EHR adoption with 2-year lag (Prob>F=0.5143)
Levels of EHR adoption	0.022	0.004	1.681	2.229	1.564	–1.547	–164.0	268.8
Average LOS	0.000	–9.46e–06	N/A[f]	N/A	0.001	–0.012	N/A	N/A
RR	N/A	N/A	0.818	2.192	N/A	N/A	–186.4	169.8
Levels of EHR adoption and average LOS	–0.000	–4.26e–06	N/A	N/A	–0.002	0.013	N/A	N/A
Levels of EHR adoption and RR	N/A	N/A	–1.663	–2.232	N/A	N/A	166.2	–271.7
Market competition (HHI[g])	0.219[h]	–0.068	0.223[h]	–0.068	3.610	–3.275	3.749	–3.103
Medicare percentage	0.063[i]	0.024	0.068[h]	0.030	–2.419	0.523	–2.416	0.783
Medicaid percentage	0.018	0.891[h]	0.018	0.092[h]	1.742	–2.822	1.741	–2.838
Years								
2014	Reference	Reference	Reference	Reference	Reference	Reference	Reference	Reference
2015	0.014[h]	–0.006	0.014[h]	–0.006	–1.057	–0.714	–1.178	–0.910
2016	0.011[h]	–0.008[i]	0.009	–0.008[i]	0.141	0.482	0.048	0.323
2017	0.010[h]	–0.018[j]	0.009[i]	–0.018[h]	0.008	0.508	–0.126	0.229

[a]OM: operating margin.

[b]LOS: length of stay.

[c]EHR: electronic health record.

[d]RR: readmission rate.

[e]TM: total margin.

[f]N/A: not applicable.

[g]HHI: Herfindahl-Hirschman Index.

[h]$P<.05$.

[i]$P<.10$.

[j]$P<.001$.

EHR: Length of Stay (Operating Margin and Total Margin)

Model 1 analyzed the relationship between EHR adoption levels and operating margins without any lags in operating margins, with LOS as a moderating variable for acute care hospitals. For Model 1, the prob>F was greater than 0.05, meaning this model did not provide a statistical explanation for the proposed relationship between EHR adoption levels and operating margins with LOS as a moderating variable.

Model 5 analyzed the relationship between EHR adoption levels and operating margins with a 1-year lag in operating margins, with LOS as the moderating variable for acute care hospitals. The prob>F was less than .05 for this model; however, the analysis did not provide statistically significant evidence to support the relationship between EHR adoption levels and operating margins with a 1-year lag in operating margins, with LOS as a moderating variable. The nonsignificant results indicated a direct positive association between EHR adoption levels and operating margins and LOS; however, when LOS acts as a moderating variable, the indirect relationship between EHR adoption levels and operating margins was negative.

Model 6 analyzed the relationship between EHR adoption levels and operating margins with a 2-year lag in operating margins, with LOS as a moderating variable for acute care hospitals. Even though the prob>F was less than .05 for this model, the analysis did not provide statistically significant evidence to support the relationship between EHR adoption levels and

operating margins with a 1-year lag in operating margins, with LOS as a moderating variable. The nonsignificant results indicated a direct positive association between EHR adoption levels and operating margins with a 2-year lag, which was expected. Additionally, the nonsignificant results indicated a direct negative association between EHR adoption levels and LOS, which is consistent with the findings from the literature. However, when LOS is introduced as a moderating variable, the nonsignificant results indicate a negative indirect relationship between EHR adoption levels and operating margins with a 2-year lag.

Model 3 analyzed the relationship between EHR adoption levels and total margins without any lags in total margins, with LOS as a moderating variable for acute care hospitals. The prob>F was greater than 0.05, meaning the models did not provide a statistically significant explanation for the proposed relationship between EHR adoption levels and total margins without any lags in total margins, with LOS as a moderating variable.

Model 9 analyzed the relationship between EHR adoption levels and total margins with a 1-year lag in total margins, with LOS as a moderating variable for acute care hospitals. For Model 9, the prob>F was greater than 0.05, meaning this model could not accurately predict the relationship between EHR adoption levels and total margins with a 1-year lag in total margins, with LOS as a moderating variable.

Model 10 analyzed the relationship between EHR adoption levels and total margins with a 2-year lag in total margins, with LOS as a moderating variable for acute care hospitals. For Model 10, the prob>F was greater than 0.05, which indicates that this model could not accurately predict the relationship between EHR adoption levels and total margins with a 2-year lag in total margins, with LOS as a moderating variable.

EHR: Readmission Rate (Operating Margin and Total Margin)

Model 2 analyzed the relationship between EHR adoption levels and operating margins without any lags in operating margins, with readmission rates as a moderating variable for acute care hospitals. Hospitals with higher readmission payment adjustment factors have lower penalties [32]. This was the only model in which the results from the analysis provided statistically significant evidence to support the proposed relationship. At the significance level of $P<.05$, EHR adoption levels were positively associated with operating margins. Similarly, at the significance level of $P<.05$, readmission rates were positively associated with an increase in operating margin. However, when readmission rates are introduced as a moderating variable, the magnitude of the relationship between levels of EHR adoption and operating margins is negative.

Model 7 analyzed the relationship between EHR adoption levels and operating margins with a 1-year lag in operating margins, with readmission rates as a moderating variable for acute care hospitals. For Model 7, the prob>F was less than .05 for this model; however, the analysis did not provide statistically significant evidence to support the relationship between EHR adoption levels and operating margins with a 1-year lag in operating margins, with readmission rates as a moderating variable. The nonsignificant results indicate a direct positive association between EHR adoption levels and operating margins with a 1-year lag and readmission rates, which is consistent with the findings from Model 2. However, when readmission rates act as a moderating variable, the nonsignificant results indicate a positive relationship between levels of EHR adoption and operating margins with a 1-year lag, which was the opposite of the results from Model 2.

Model 8 analyzed the relationship between EHR adoption levels and operating margins with a 2-year lag in operating margins, with readmission rates as a moderating variable for acute care hospitals. The prob>F was less than .05 for this model; however, the analysis did not provide statistically significant evidence to support the relationship between EHR adoption levels and operating margins with a 2-year lag in operating margins, with readmission rates as a moderating variable. Similar to Model 7, the nonsignificant results indicated a direct positive association between EHR adoption levels and operating margins with a 2-year lag and readmission rates, which was consistent with the findings from Model 2. However, when readmission rates act as a moderating variable, the nonsignificant results indicated a positive relationship between levels of EHR adoption and operating margins with a 2-year lag, which was the opposite of the results from Model 2.

Model 4 analyzed the relationship between EHR adoption levels and total margins without any lags in total margins, with readmission rates as a moderating variable for acute care hospitals. The prob>F was greater than 0.05, meaning this model could not provide a statistically significant explanation for the proposed relationship between EHR adoption levels and total margins without any lags in total margins, with readmission rates as a moderating variable.

Model 11 analyzed the relationship between EHR adoption levels and total margins with a 1-year lag in total margins, with readmission rates as a moderating variable for acute care hospitals. For Model 11, the prob>F was greater than 0.05, which indicates that this model could not accurately predict the relationship between EHR adoption levels and total margins with a 1-year lag in total margins, with readmission rates as a moderating variable.

Model 12 analyzed the relationship between EHR adoption levels and total margins with a 2-year lag in total margins, with readmission rates as a moderating variable for acute care hospitals. The prob>F was greater than 0.05, meaning this model could not provide a statistically significant explanation for the proposed relationship between EHR adoption levels and total margins with a 2-year lag in total margins, with readmission rates as a moderating variable.

Results from the regression analysis with fixed effects are displayed in Tables 3 and 4. Table 3 includes results from the regression analysis with financial and clinical outcomes from the same year. Hospitals receive their reimbursement and penalties associated with readmission rates and LOS approximately 1 to 2 years after the actual outcomes occur. In order to accommodate this situation, operating margin and total margin ratios were calculated with a 1- and 2-year lag. Table 4

presents results with lags in profit margins for acute care hospitals.

The results from Table 3 for model 2 suggest that, at the significance level of $P<.05$, a 1-unit increase in EHR adoption was associated with an increase of approximately 5.34% in the operating margin.

Table 4 displays results from the analyses with the added lag effect in operating and total margins. According to the results displayed in Table 4, it can be inferred that at the significance levels of $P<.05$, $P<.10$, or $P<.001$, there is not enough evidence to support models 5-8 from this study. For models 9-12, the models did not provide a statistical explanation for the proposed relationships. In other words, the models discussed above could not accurately predict the proposed relationships.

Discussion

Overview

The objective of this study was to determine how EHR adoption level contributes to financial and clinical outcomes for acute care hospitals.

To understand the relationship between EHR adoption level and financial outcomes, moderated by clinical outcomes, this study used a fixed effects moderation analysis model. We hypothesized that there would be a positive association between EHR adoption level and operating and total margins, with LOS and readmission rates as moderating variables.

According to the results displayed in Table 3, for models 1, 3, and 4, the prob>F was greater than .05, meaning the models did not provide a statistical explanation for the proposed relationships in these models. In other words, the models discussed above could not accurately predict the proposed relationships, and there is no evidence that EHR adoption levels have a linear relationship with or explain variance in the operating margins, total margins, and LOS.

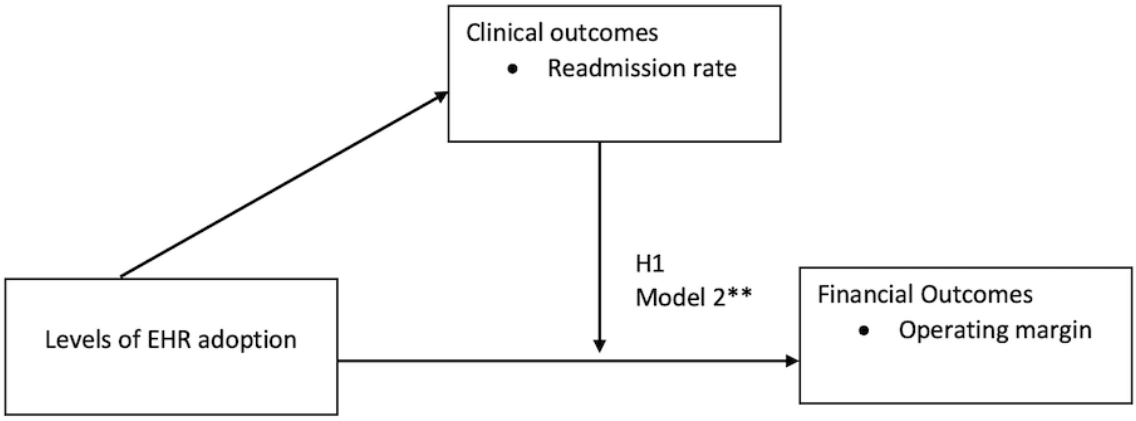
Even though the results are inverse of what was predicted in the hypothesis, these findings indicated that the relationship between EHR adoption levels and operating margins was statistically significant when it was moderated by the readmission rates variable at the significance level of $P<.05$. According to the results of the moderation analysis, as the readmission rate increases by 1 unit, the effect of a 1-unit increase in EHR adoption level on the operating margin decreases by 5.38%. In other words, when the hospital incurred lower penalties for readmissions, the operating margins increased. The minimum payment adjustment factor is 0.97 (ie, 3% maximum penalty). The maximum payment adjustment factor is 1 (ie, no penalty), and hospitals with higher payment adjustment factors have lower penalties and, in turn, larger operating margins [32].

In order to confirm any lagged effect (the timeline of hospitals receiving penalties or incentives for EHR adoption being not clear), this study included additional models that accounted for 1- and 2-year lag in the profit margin ratios (models 5-12). The results, however, did not provide any statistically significant evidence supporting a positive relationship between EHR adoption level and profit margin ratios when the lag effect was included in the model.

Findings from current literature indicate an improvement in LOS as a result of EHR adoption (not necessarily adoption level) yielding increased compensation for the loss of patient days from Center for Medicare and Medicaid Services [25]; however, for this study, none of the tested models provided a statistical explanation for the proposed relationships between EHR adoption and profit margins with LOS as moderating variables.

Even though this finding is opposite of what was proposed in the hypothesis, this finding provides statistically significant evidence that levels of EHR adoption change operating margins when readmission rates are taken into account (Figure 3). Analyzing more recent data could indicate a decrease in readmission rates as a result of increased levels of EHR adoption, yielding an increase in operating margins. The relationship between EHR adoption level and operating margins has not been previously evaluated using readmission rates as moderating variables. Hence, this finding from this study is a unique contribution to the current literature.

Figure 3. Electronic health record (EHR) value analysis framework with results. ***P*<.05.



Limitations of This Study

Regardless of the valuable contribution of the buildout of the conceptual model and the results from the analysis, this study has limitations. First, there is always a risk when using secondary data to conduct research that was not the intent when the data were collected, as this could result in inconsistency in the data collection methods due to the possibility of human error [36].

Second, this study used data from the Medicare Cost Reports to operationalize the readmission rate variable. This particular measure is reported on a 3-year rolling basis, meaning the data analyzed included a rolling average of 3 years of readmission rate data for each hospital [32]. This study operationalized the readmission rate data for specific years in order to evaluate their relationship with levels of EHR adoption and financial outcomes, which can be considered a limitation.

Conclusion

The HITECH Act has played an important role in EHRs becoming an integral part of the modern health system over the last 10 years. The goal of enacting the HITECH Act of 2009 was to reduce health care costs, improve the quality of the care provided, and increase patient safety for providers and organizations that exhibited meaningful use of certified EHR systems [1,37]. Given the cost and complexity of EHR adoption, analyzing its value from various and seemingly atypical perspectives is essential.

The current literature does a good job of providing perspectives on EHR value relative to individual financial and clinical outcomes, but it falls short in providing a collective value analysis. This study fills the gap in the literature by evaluating individual relationships between EHR adoption levels and financial and clinical outcomes, in addition to evaluating the relationship between EHR adoption level and financial outcomes, with clinical outcomes as moderators.

This study provided statistically significant evidence, indicating that there is a relationship between EHR adoption level and operating margins when this relationship is moderated by readmission rates. This finding could further be supported by evaluating more recent data to analyze whether hospitals increasing their level of EHR adoption would decrease readmission rates, resulting in an increase in operating margins. Hospitals would incur lower penalties as a result of improved readmission rates, which would contribute toward improved operating margins.

Conflicts of Interest

Not applicable.

References

1. McAlearney AS, Sieck C, Hefner J, Robbins J, Huerta TR. Facilitating ambulatory electronic health record system implementation: evidence from a qualitative study. Biomed Res Int 2013;2013:629574 [FREE Full text] [doi: 10.1155/2013/629574] [Medline: 24228257]
2. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
3. Adler-Milstein J, Green CE, Bates DW. A survey analysis suggests that electronic health records will yield revenue gains for some practices and losses for many. Health Aff (Millwood) 2013;32(3):562-570. [doi: 10.1377/hlthaff.2012.0306] [Medline: 23459736]
4. Collum TH, Menachemi N, Sen B. Does electronic health record use improve hospital financial performance? Evidence from panel data. Health Care Manage Rev 2016;41(3):267-274 [FREE Full text] [doi: 10.1097/HMR.0000000000000068] [Medline: 26052785]
5. Jang Y, Lortie MA, Sanche S. Return on investment in electronic health records in primary care practices: a mixed-methods study. JMIR Med Inform 2014 Oct 29;2(2):e25 [FREE Full text] [doi: 10.2196/medinform.3631] [Medline: 25600508]
6. Pine R, Tart K. Return on investment: benefits and challenges of baccalaureate nurse residency program. Nurs Econ 2007;25(1):13-18. [Medline: 17402673]
7. Payne TH, Bates DW, Berner ES, Bernstam EV, Covvey HD, Frisse ME, et al. Healthcare information technology and economics. J Am Med Inform Assoc 2013;20(2):212-217 [FREE Full text] [doi: 10.1136/amiajnl-2012-000821] [Medline: 22781191]
8. Peterson LT, Ford EW, Eberhardt J, Huerta TR, Menachemi N. Assessing differences between physicians' realized and anticipated gains from electronic health record adoption. J Med Syst 2011 May;35(2):151-161 [FREE Full text] [doi: 10.1007/s10916-009-9352-z] [Medline: 20703574]
9. Rudin RS, Friedberg MW, Shekelle P, Shah N, Bates DW. Getting value from electronic health records: research needed to improve practice. Ann Intern Med 2020 Jul 02;172(11 Suppl):S130-S136 [FREE Full text] [doi: 10.7326/M19-0878] [Medline: 32479182]
10. Shah GH, Leider JP, Castrucci BC, Williams KS, Luo H. Characteristics of local health departments associated with implementation of electronic health records and other informatics systems. Public Health Rep 2016;131(2):272-282 [FREE Full text] [doi: 10.1177/003335491613100211] [Medline: 26957662]

11. Feldman SS. Value proposition of health information exchange. In: Dixon BE, editor. Health Information Exchange: Navigating and Managing a Network of Health Information Systems. Amsterdam, Netherlands: Academic Press; 2015.
12. Feldman SS. Public-Private Interorganizational Sharing of Health Data for Disability Determination. Ann Arbor, MI: ProQuest LLC; 2011.
13. Copeland TE, Weston JF, Shastri K. Financial Theory and Corporate Policy. Harlow, UK: Pearson; 2014.
14. Thouin MF, Hoffman JJ, Ford EW. The effect of information technology investment on firm-level performance in the health care industry. Health Care Manage Rev 2008;33(1):60-68 [FREE Full text] [doi: 10.1097/01.HMR.0000304491.03147.06] [Medline: 18091445]
15. Edwardson N, Kash BA, Janakiraman R. Measuring the impact of electronic health record adoption on charge capture. Med Care Res Rev 2017 Oct;74(5):582-594 [FREE Full text] [doi: 10.1177/1077558716659408] [Medline: 27416948]
16. Lim MC, Boland MV, McCannel CA, Saini A, Chiang MF, Epley KD, et al. Adoption of electronic health records and perceptions of financial and clinical outcomes among ophthalmologists in the United States. JAMA Ophthalmol 2018 Mar 01;136(2):164-170 [FREE Full text] [doi: 10.1001/jamaophthalmol.2017.5978] [Medline: 29285542]
17. Fleming NS, Becker ER, Culler SD, Cheng D, McCorkle R, da Graca B, et al. The impact of electronic health records on workflow and financial measures in primary care practices. Health Serv Res 2014 Mar;49(1 Pt 2):405-420 [FREE Full text] [doi: 10.1111/1475-6773.12133] [Medline: 24359533]
18. Wang T, Wang Y, McLeod A. Do health information technology investments impact hospital financial performance and productivity? International Journal of Accounting Information Systems 2018 Mar;28:1-13 [FREE Full text] [doi: 10.1016/j.accinf.2017.12.002]
19. Knepper MM, Castillo EM, Chan TC, Guss DA. The effect of access to electronic health records on throughput efficiency and imaging utilization in the emergency department. Health Serv Res 2018 Apr;53(2):787-802 [FREE Full text] [doi: 10.1111/1475-6773.12695] [Medline: 28376563]
20. Litzelman DK, Dittus RS, Miller ME, Tierney WM. Requiring physicians to respond to computerized reminders improves their compliance with preventive care protocols. J Gen Intern Med 1993 Jul;8(6):311-317 [FREE Full text] [doi: 10.1007/BF02600144] [Medline: 8320575]
21. 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 [FREE Full text] [doi: 10.1001/archinternmed.2008.520] [Medline: 19171805]
22. Daniel GW, Ewen E, Willey VJ, Reese Iv CL, Shirazi F, Malone DC. Efficiency and economic benefits of a payer-based electronic health record in an emergency department. Acad Emerg Med 2010 Aug;17(8):824-833 [FREE Full text] [doi: 10.1111/j.1553-2712.2010.00816.x] [Medline: 20670319]
23. Everson J, Lee SYD, Friedman CP. Reliability and validity of the American Hospital Association's national longitudinal survey of health information technology adoption. J Am Med Inform Assoc 2014 Oct;21(e2):e257-e263 [FREE Full text] [doi: 10.1136/amiajnl-2013-002449] [Medline: 24623194]
24. Adler-Milstein J, Everson J, Lee SYD. EHR adoption and hospital performance: time-related effects. Health Serv Res 2015 Dec;50(6):1751-1771 [FREE Full text] [doi: 10.1111/1475-6773.12406] [Medline: 26473506]
25. Mirani R, Harpalani A. Business benefits or incentive maximization? impacts of the medicare EHR incentive program at acute care hospitals. ACM Trans Manage Inf Syst 2013 Dec;4(4):1-19 [FREE Full text] [doi: 10.1145/2543900]
26. Thirukumaran CP, Dolan JG, Reagan Webster P, Panzer RJ, Friedman B. The impact of electronic health record implementation and use on performance of the Surgical Care Improvement Project measures. Health Serv Res 2015;50(1):273-289. [doi: 10.1111/1475-6773.12191] [Medline: 24965357]
27. Wani D, Malhotra M. Does the meaningful use of electronic health records improve patient outcomes? J Oper Manag 2018;60(1):1-18. [doi: 10.1016/j.jom.2018.06.003]
28. Medicare and medicaid promoting interoperability program basics. Centers for Medicare & Medicaid Services. 2018. URL: https://www.cms.gov/medicare/regulations-guidance/promoting-interoperability-programs/medicare-medicaid-basics [accessed 2019-01-16]
29. Rojas-García A, Turner S, Pizzo E, Hudson E, Thomas J, Raine R. Impact and experiences of delayed discharge: A mixed-studies systematic review. Health Expect 2018 Mar;21(1):41-56 [FREE Full text] [doi: 10.1111/hex.12619] [Medline: 28898930]
30. Pink GH, Holmes GM, D'Alpe C, Strunk LA, McGee P, Slifkin RT. Financial indicators for critical access hospitals. J Rural Health 2006;22(3):229-236 [FREE Full text] [doi: 10.1111/j.1748-0361.2006.00037.x] [Medline: 16824167]
31. Schreiber R, Shaha SH. Computerised provider order entry adoption rates favourably impact length of stay. J Innov Health Inform 2016 May 18;23(1):166 [FREE Full text] [doi: 10.14236/jhi.v23i1.166] [Medline: 27348485]
32. Hospital readmissions reduction program (HRRP). Centers for Medicare & Medicaid Services. 2020. URL: https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program [accessed 2023-11-08]
33. Ben-Assuli O, Shabtai I, Leshno M. The impact of EHR and HIE on reducing avoidable admissions: controlling main differential diagnoses. BMC Med Inform Decis Mak 2013 May 17;13(1):49 [FREE Full text] [doi: 10.1186/1472-6947-13-49] [Medline: 23594488]

34. Lee J, Kuo YF, Lin YL, Goodwin JS. The combined effect of the electronic health record and hospitalist care on length of stay. Am J Manag Care 2015 Mar 01;21(3):e215-e221 [FREE Full text] [Medline: 26014309]
35. Bailey MA. Real Econometrics The Right Tools to Answer Important Questions. New York, NY: Oxford University Press; 2017.
36. Hoffmann F, Andersohn F, Giersiepen K, Scharnetzky E, Garbe E. [Validation of secondary data. Strengths and limitations]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2008 Oct;51(10):1118-1126. [doi: 10.1007/s00103-008-0646-y] [Medline: 18985405]
37. Redd TK, Read-Brown S, Choi D, Yackel TR, Tu DC, Chiang MF. Electronic health record impact on productivity and efficiency in an academic pediatric ophthalmology practice. J AAPOS 2014 Dec;18(6):584-589 [FREE Full text] [doi: 10.1016/j.jaapos.2014.08.002] [Medline: 25456030]

Abbreviations

AHA: American Hospital Association
EHR: electronic health record
HHI: Herfindahl-Hirschman Index
HITECH: Health Information Technology for Economic and Clinical Health
LOS: length of stay

Edited by J Hefner; submitted 06.09.23; peer-reviewed by L Heryawan, A Kotlo; comments to author 23.10.23; revised version received 29.10.23; accepted 29.11.23; published 24.01.24.

Please cite as:
Modi S, Feldman SS, Berner ES, Schooley B, Johnston A
Value of Electronic Health Records Measured Using Financial and Clinical Outcomes: Quantitative Study
JMIR Med Inform 2024;12:e52524
URL: https://medinform.jmir.org/2024/1/e52524
doi:10.2196/52524
PMID:38265848

Review

Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review

Dobin Yim[1], PhD; Jiban Khuntia[2], PhD; Vijaya Parameswaran[3], RD, PhD; Arlen Meyers[2], PhD

[1]Loyola University, Maryland, MD, United States

[2]University of Colorado Denver, Denver, CO, United States

[3]Stanford University, Stanford, CA, United States

Corresponding Author:
Jiban Khuntia, PhD
University of Colorado Denver
1475 Lawrence St.
Denver, CO
United States
Phone: 1 3038548024
Email: jiban.khuntia@ucdenver.edu

Abstract

Background: Generative artificial intelligence tools and applications (GenAI) are being increasingly used in health care. Physicians, specialists, and other providers have started primarily using GenAI as an aid or tool to gather knowledge, provide information, train, or generate suggestive dialogue between physicians and patients or between physicians and patients' families or friends. However, unless the use of GenAI is oriented to be helpful in clinical service encounters that can improve the accuracy of diagnosis, treatment, and patient outcomes, the expected potential will not be achieved. As adoption continues, it is essential to validate the effectiveness of the infusion of GenAI as an intelligent technology in service encounters to understand the gap in actual clinical service use of GenAI.

Objective: This study synthesizes preliminary evidence on how GenAI assists, guides, and automates clinical service rendering and encounters in health care The review scope was limited to articles published in peer-reviewed medical journals.

Methods: We screened and selected 0.38% (161/42,459) of articles published between January 1, 2020, and May 31, 2023, identified from PubMed. We followed the protocols outlined in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to select highly relevant studies with at least 1 element on clinical use, evaluation, and validation to provide evidence of GenAI use in clinical services. The articles were classified based on their relevance to clinical service functions or activities using the descriptive and analytical information presented in the articles.

Results: Of 161 articles, 141 (87.6%) reported using GenAI to assist services through knowledge access, collation, and filtering. GenAI was used for disease detection (19/161, 11.8%), diagnosis (14/161, 8.7%), and screening processes (12/161, 7.5%) in the areas of radiology (17/161, 10.6%), cardiology (12/161, 7.5%), gastrointestinal medicine (4/161, 2.5%), and diabetes (6/161, 3.7%). The literature synthesis in this study suggests that GenAI is mainly used for diagnostic processes, improvement of diagnosis accuracy, and screening and diagnostic purposes using knowledge access. Although this solves the problem of knowledge access and may improve diagnostic accuracy, it is oriented toward higher value creation in health care.

Conclusions: GenAI informs rather than assisting or automating clinical service functions in health care. There is potential in clinical service, but it has yet to be actualized for GenAI. More clinical service–level evidence that GenAI is used to streamline some functions or provides more automated help than only information retrieval is needed. To transform health care as purported, more studies related to GenAI applications must automate and guide human-performed services and keep up with the optimism that forward-thinking health care organizations will take advantage of GenAI.

(JMIR Med Inform 2024;12:e52073) doi:10.2196/52073

KEYWORDS

generative artificial intelligence tools and applications; GenAI; service; clinical; health care; transformation; digital

Introduction

Background

Generative artificial intelligence tools and applications (GenAI) systems automatically learn patterns and structures from text, images, sounds, animation, models, or other media inputs to generate new data with similar characteristics [1]. GenAI is used to search, write, and create models, computer codes, and art forms without human assistance. GenAI has emerged significantly in the current decade to help every industry through different products such as ChatGPT, Bing Chat, Bard, LLaMA, Stable Diffusion, Midjourney, and DALL-E [2-5]. Almost all industries share an optimistic vision, with significant investment in using GenAI to transform aspects of value chains [6-10]. However, similar to many other technology hypes, whether this optimism will translate to value outcomes or be a "fad or fashion" remains to be tested over time.

The adoption of GenAI in health care is emerging. Studies point to the use of GenAI in service interactions involving breast cancer diagnoses [11], bariatric surgery [12], cardiopulmonary resuscitation [13], and breast cancer radiologic decision-making [14]. GenAI has the potential to transform by performing tasks at higher quality than humans, which may reduce errors associated with humans in expert domains such as cancer detection [15] and neurological clinical decisions [16]. The rise of GenAI is also referred to as the "second machine age" [17], whereby "instead of machines performing mechanical work they are taking on cognitive work exclusively in the human domain" [17]. Although these instances are encouraging, how exactly GenAI helps in health care processes needs to be articulated and evaluated to provide an understanding of use and value linkages [18,19]. Thus, we asked the following research questions (RQs) in this study: (1) How is GenAI used across different aspects of health care services? (RQ 1) and (2) What is the preliminary evidence of GenAI use across health care services? (RQ 2).

It is essential to explore these 2 RQs for several reasons. Exploring GenAI's use in health care services is essential for realizing its potential benefits, addressing ethical concerns, and continually improving its applications to enhance patient care and the health care ecosystem. This impact spans different areas. For instance, GenAI can help analyze data to provide personalized treatment and tailor interventions. It has shown promise in improving diagnostic accuracy, with higher levels of accuracy in the interpretation of images and scans. AI applications can enhance patient engagement by providing personalized health recommendations, reminders for medications, and real-time monitoring of vital signs. On the provider side, GenAI can save costs by streamlining administrative tasks and improving efficiency, early disease detection, and preventive care. Similarly, knowing the preliminary evidence of GenAI use across health care services is crucial for making informed decisions, ensuring regulatory compliance, building trust, guiding research initiatives, and addressing ethical considerations. This sets the stage for the responsible and effective integration of GenAI into the health care landscape.

The impact of GenAI in health care depends on various factors, including the specific application, quality of data used for training, ethical considerations, and regulatory framework in place. Continuous monitoring, evaluation, and responsible deployment are essential to maximize the positive impact and mitigate potential negative consequences. For instance, artificial intelligence (AI) assists pathologists in diagnosing diseases from pathology slides, leading to faster and more accurate diagnoses and improving patient outcomes [20]. Analysis of oncology literature, clinical trial data, and patient records can help oncologists identify personalized, evidence-based treatment options for patients with cancer, potentially improving treatment decisions [21]. AI has been applied to analyze medical images for conditions such as diabetic retinopathy, aiding in early detection and intervention [22]. AI analyzes clinical and molecular data to help physicians make more informed decisions about cancer treatment and steer them toward personalized and effective therapies [23].

Concerns about using GenAI remain because of algorithmic bias in predictive models that causes discrimination, unequal distribution of health care resources, and exacerbated health disparities [24]. Data privacy and the need for clear guidelines on AI in health care remain a gap, with reported misuse [25]. Misinterpretations or errors in algorithms can lead to incorrect diagnoses, specifically for image readings, which underscores the importance of human oversight in critical health care decisions [26]. Furthermore, implementing and maintaining AI systems can be costly, and overreliance on technology without sufficient human oversight may result in overlooking critical clinical nuances and potentially compromising patient care [27]. Therefore, it is essential to note that the impact of AI on health care is a dynamic and evolving field. Regular updates and scrutiny of the latest research and applications are necessary to understand the positive and negative aspects of GenAI in health care.

Using a literature scoping, review, and synthesis approach in this study, we evaluated the proportionate evidence of using GenAI to assist, guide, and automate clinical service functions. Technologies in general help standardize [28], provide flexibility [29], increase experience and satisfaction through relational benefits [30], induce higher switching costs [31], and enhance the overall quality [32] and value [33] of services. However, high technology may reduce personal touch, trust, and loyalty in service settings [34-38]. Complex technologies may introduce anxiety, confusion, and isolation [39] or disconnection, disruption, and passivity stressors [13] that can erode satisfaction, loyalty, and retention in service settings [28,40-42]. Given the mixed evidence in previous research on the role of technology in services [28,43,44], it is timely to assess to what extent GenAI may even have a role in shaping or disrupting health care services. Overall, the ground realities of the potential for emerging GenAI to benefit health care services rather than just being another knowledge and collation tool need to be assessed and reported to influence further research and practice activities.

Objectives

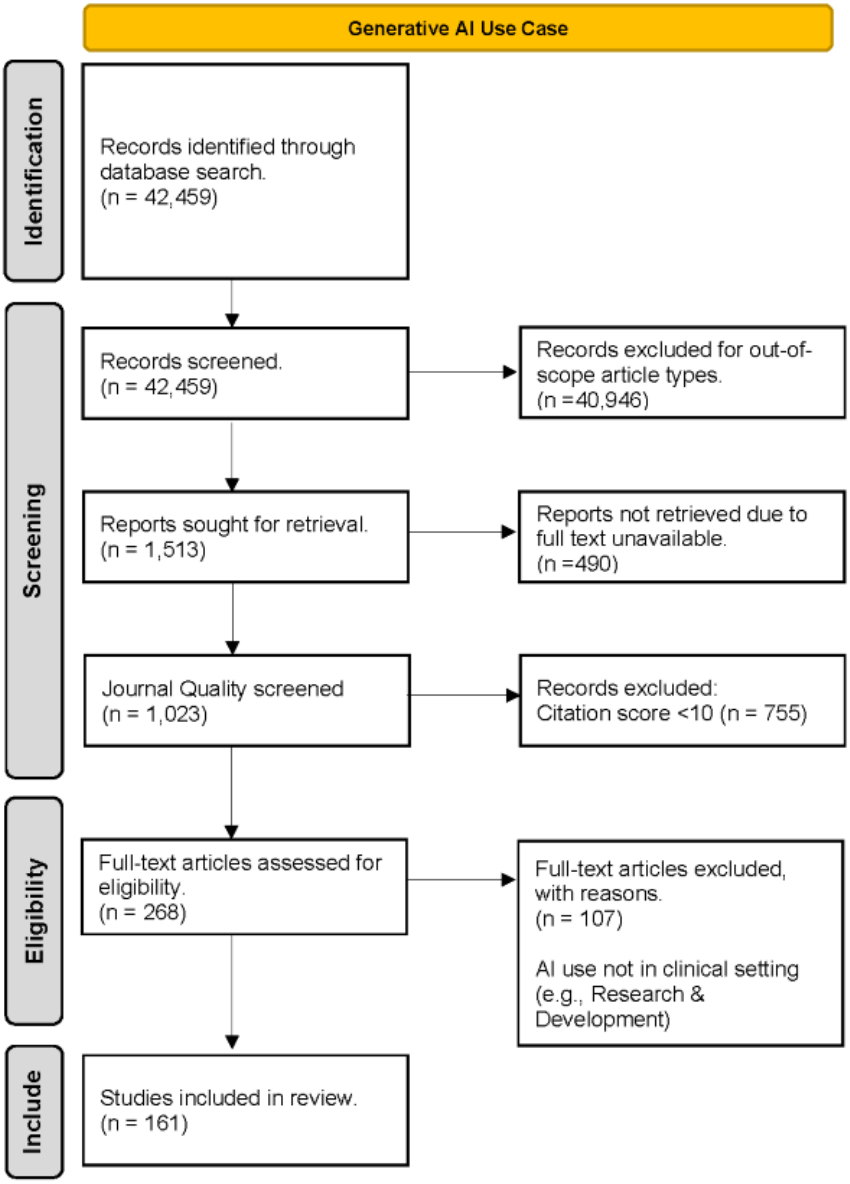
This study took a deep dive to review and synthesize preliminary evidence on how GenAI is used to assist, guide, and automate activities or functions during clinical service encounters in health care, with plausible indications for differential use. More evidence on the actual use is needed to assert that GenAI plays a considerable role in the digital transformation of health care. Therefore, this study aims to identify how GenAI is used in clinical settings by systematically reviewing preliminary evidence on its applications to assist, guide, and automate clinical activities or functions.

Methods

Article Search and Selection Strategy

This study aims to identify how physicians use GenAI in clinical settings, as evidenced in published studies. The design of this study adheres to the protocols outlined in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [45,46]. Figure 1 provides a flowchart of this study's article search and inclusion process.

Figure 1. Literature screening process for relevant articles on generative artificial intelligence (AI) tools and applications.



We focused our search exclusively on PubMed to ensure the credibility of this study's medical or clinical service settings. PubMed is part of the National Library of Medicine and a trusted national source of peer-reviewed publications on medical devices, software applications, and techniques used in the clinical setting. We performed keyword searches to retrieve relevant GenAI publications in PubMed that used "artificial intelligence" anywhere in the text of the article written in English. The sampling period of the publications was from January 1, 2020, to May 31, 2023. The search yielded 42,459 results in the first round of identification of articles for evaluation.

Within PubMed's classification system for articles, we used the "article type" that described the material presented in the article (eg, review, clinical trial, retracted publication, or letter). We

used this article type feature in the PubMed classification system to identify peer-reviewed articles and other relevant types of publications that are pertinent to our study. A total of 52.02% (22,086/42,459) of the returned articles did not have an article type assigned from the 75 article types in PubMed's classification system and were excluded from the study sample. We included clinical, multicenter, case report, news, evaluation, and validation studies. We excluded article types that were out of scope, such as uncategorized articles, government-funded studies, reviews, editorials, errata, opinion articles, nonscientific articles, retracted publications, and supplementary files. We also excluded preprint article types that were unlikely to have attracted attention. Errata or retracted publications (404/42,459, 0.95%), supplementary files (117/42,459, 0.28%), and 50 article types that had too few search returns (243/42,459, 0.57%) were also excluded.

The screening stage excluded review articles (6732/42,459, 15.86%) with an objective that was neither aligned with nor redundant to this study's goal. Opinion articles such as editorials, letters, and commentaries were excluded (2455/42,459, 5.78%). Articles whose funding came from the government or a government agency were not considered because of a conflict of interest for the researchers of the evaluated study (8936/42,459, 21.05%), and preprint articles (77/42,459, 0.2%) were excluded because of lack of availability to the public. We also considered the full text availability of the article, and 32.39% (490/1513) of the articles were excluded in the eligibility stage.

The resulting set of records included 1023 publications. To ensure the credibility of the publication source, we used CiteScore (Elsevier) [47] as a citation index to remove publication sources whose influence is limited. Any publication source whose citation index was unavailable or <10 was removed, resulting in 268 records.

In total, 2 raters, 1 author (DY) and 1 graduate assistant (BB), evaluated 161 articles. The 2 raters' agreement was 91.93%, and the expected agreement was 82.99%. The κ score was 0.5252 (SE 0.0544; *Z* score=9.66; probability>Z score=0.0000). The author and the graduate student performed manual coding by reading the paper's title, abstract, and introduction paragraph to gain a preliminary understanding of the study. After reading the abstract and introduction paragraph, each rater classified each article according to the definition of the 3 classes. For articles that were difficult to understand, the rater read the article further to gain a better understanding of the article. We defined clinical service settings to include the life cycle of physician encounters with patients for the diagnosis, prognosis, and management of health conditions. The research and development of drug discovery, for instance, was not considered. This process eliminated 107 records. The final data set of articles considered for this study was 161.

Ethical Considerations

The data collected for this study were obtained from publicly available sources. The study did not involve any interaction with users. Therefore, ethics approval was not required for this study.

Data Extraction and Categorization Process

We adopted a modified thematic synthesis approach for data analysis that involved coding the text, developing descriptive themes, and generating analytical themes [48]. Initially, each author coded each line of text extracted from the articles, assigning it to different dimensions. This line-by-line coding process facilitated identifying and capturing critical article information and concepts. Next, each author developed descriptive themes by grouping related codes and identifying common patterns or topics emerging from the coded data. These descriptive themes provided a broad overview of the various aspects of AI in the clinical service context. Building on the descriptive themes, each author generated analytical pieces to deepen the understanding and interpretation of the data. The analytical themes involved exploring relationships, connections, and implications within and across the articles, allowing for the extraction of meaningful insights.

Throughout the analysis process, all the authors engaged in extensive discussions to refine and finalize the results of the thematic synthesis. By collectively examining and interpreting the data, the research team ensured the robustness and reliability of the synthesized findings. Similar dimensions were then merged to generate the following 3 meaningful dimensions (assist, guide, and automate) and for relevance to the study objectives, as shown in Textbox 1. The researchers manually coded each article into several groups. They then tried to synthesize them into 1 of the 3 categories of *assist*, *guide*, and *automate* by looking at the title, abstract, and introduction (where applicable).

Textbox 1. Use of generative artificial intelligence tools and applications in clinical services in the reviewed articles (N=161).

Assist

- Improve diagnostic accuracy or reduce error by accessing knowledge during clinical services (141/161, 87.6%) [49-96]
- Activities:
  - Disease detection (19/161, 11.8%) [58,63,67,69,71,73,77,90,97-107]
  - Diagnosis (14/161, 8.7%) [100,108-120]
  - Screening (12/161, 7.5%) [65,86,87,93,121-128]
- Service areas:
  - Radiology (17/161, 10.6%) [49-63,65,66]
  - Cardiology (12/161, 7.5%) [67-72,74,76-79,129]
  - Gastrointestinal medicine (4/161, 2.5%) [81-84]
  - Diabetes (6/161, 3.7%) [86-91]
- Approaches and methods:
  - Deep learning (34/161, 21.1%) [49,59,60,62,63,65,68,71,79,89,100,107,108,111,115,123,125,130-145]
  - Machine learning (9/161, 5.6%) [53,55,83,91,110,146-149]
  - Image analysis (13/161, 8.1%) [68,88,104,110,111,114,116,119,133,135,138,150,151]

Guide

- Recommend treatment options, step-by-step instructions, or checklists to improve clinical services (13/161, 8.1%) [64,80,85,96,152-160]
- Personalized treatment plans (1/161, 0.6%) [64]
- Monitoring and managing (1/161, 0.6%) [96]

Automate

- Minimize or eliminate human provider involvement in clinical services or follow-ups (7/161, 4.3%) [94,95,161-165]

In addition to manual coding by human researchers, we used ChatGPT (version 3.5; OpenAI) for automatic coding. ChatGPT-3.5 was used for speed and cost. ChatGPT-4 is less accessible to users who do not have the funds to pay for its monthly subscription. ChatGPT-3.5 training used one-shot learning using the standard user interface with the "foundational" mode, and no fine-tuning was performed. Future studies may use focused data sets for fine-tuning to improve classification accuracy. However, our study demonstrates that classification accuracy is high and robust even without fine-tuning. This procedure was implemented to check for any subjective bias and demonstrate AI's potential use to complement the human coding process. The abstracts and introductions of these 161 articles were fed into ChatGPT using in-context or a few short learning processes that fine-tune a pair of domain-specific inputs and outputs to train, thereby enhancing the relevance and accuracy of ChatGPT's automated coding output [166,167].

For instance, a sample of input we used in the study was the abstract, which summarizes the article. The output is the categories identified by the experts. ChatGPT learns how to code a set of articles by repeating the pair of inputs and outputs. One-shot learning, which consists of a single pair of inputs and outputs in general, performs as well as >2 samples and zero-shot learning. The benefits of in-context learning (ICL) in ChatGPT include enhanced relevance, where the foundational model becomes better at generating content for domain-specific tasks without additional training of the full model; controlled output such as developing a single word matching the desired coding category or variable; and reduced biases inherent in manual coding. We used the definitions provided in Textbox 1 to train and restrict ChatGPT to choose only 1 of the 3 use-case categories. We further compared ChatGPT's classification with expert coding and found a high level of agreement between the 2, with a κ score of 0.94.

As mentioned previously, the manual coding process involved the raters coding and evaluating each article. After each rater coded the article, the results were compared and discussed to further refine the classification definition and derive consensus on the final assignment of the article classification. This "gold standard" classification was compared with automatic coding performed by ChatGPT (version 3.5). Automatic coding was performed by ChatGPT-3.5. Classification training was performed using one-shot ICL. ChatGPT learns how to classify articles by being fed a pair of articles and classification labels. For example, a user can feed a prompt or use control tokens to indicate an article abstract and the label associated with the article. In our context, 3 articles and labels were fed to the interface. After this initial prompt session of training on 3 classification labels, subsequent interactions of providing only

the article abstract with a prompt asking for a class label would return ChatGPT's prompt completion. Alternatively, training could involve >1 example of the article and its label, which would then be called *few-shot learning*. To summarize, 161 articles were coded by ChatGPT-3.5 based on a single instance of ICL.

Results

Findings From the Synthesis on the Use of GenAI to Assist in Different Aspects of Health Care Services

GenAI can improve clinical services in 3 ways. First, of the 161 articles, 141 (87.6%) reported using GenAI to assist services through knowledge access, collation, and filtering. The assistance of GenAI was used for disease detection (19/161, 11.8%) [58,63,67,69,71,73,77,90,97-107], diagnosis (14/161, 8.7%) [100,108-120], and screening processes (12/161, 7.5%) [65,86,87,93,121-127,168,169] in the areas of radiology (17/161, 10.6%) [49-63,65,66], cardiology (12/161, 7.5%) [67-72,74,76-79,129], gastrointestinal medicine (4/161, 2.5%) [81-84], and diabetes (6/161, 3.7%) [86-91]. Thus, although the use of GenAI has percolated across almost all disease-relevant and main service–relevant areas in health care, it is mainly for assisting through knowledge access, collation, and filtering.

The use of GenAI in disease diagnosis has long-term implications. For instance, identifying "referrable" diabetic retinopathy using routinely collected data would help in population health planning and prevention [86-90]; however, rigorous testing and validation of the applications are critical before clinical implementation [94]. Similarly, using GenAI in remote care helps improve glycemia and weight loss [95], yet challenges related to variable patient uptake and increased clinician participation necessitated by shared decision-making must be considered [96]. In radiology services, prediction models using deep learning and machine learning methods for predictive accuracy and as diagnostic aids have shown potential, and natural language processing has been used to improve readability by generating captions; however, studies report using high-quality images, highlighting the need for a future standardized pipeline for data collection and imaging detection.

In cardiology, AI analysis allows for early detection, population-level screening, and automated evaluation. It expands the reach of electrocardiography to clinical settings in which immediate interrogation of anatomy and cardiac function is needed and to locations with limited resources [67-69,71,73-75,95]. Nevertheless, there is evidence suggesting that integrating AI with patient data, including social determinants of health, enables disease prediction and early disease identification, which could lead to more precise and timely diagnoses, improving patient outcomes.

GenAI aids in diagnostic accuracy, although its focus on higher value creation in health care is limited. The articles in this review reported that they used deep learning (34/161, 21.1%) [49,59,60,62,63,65,68,71,79,89,100,107,108,111,115,123,125,130-145], machine learning (9/161, 5.6%) [53,55,83,91,110,146-149], and image analysis approaches of GenAI during the assistance process (13/161, 8.1%) [68,88,104,110,111,114,116,119,133,135,138,150,151]. Knowledge access using GenAI has the potential to enable more options and flexibility in serving patients.

Evidence of GenAI Use for Guiding or Automation Services

Only 8.1% (13/161) of the studies provided insights into how GenAI is used to guide some services by seeking recommended treatment options, step-by-step instructions, or checklists to improve clinical services [64,80,85,96,152-160]. Of the 161 studies, 1 (0.6%) study sought personalized treatment plans and discussed monitored and managed service processes using GenAI [96]. Although this use category is nascent, GenAI can help provide speed efficiency and customized solutions in health services as in other contexts [37,127,170].

Finally, only 4.3% (7/161) of the articles indicated the use of GenAI to automate any service functions that could minimize or eliminate human provider involvement. When used appropriately, automation provides a predictable, reliable, and faster experience everywhere, every time for all customers, which will be a standardized way to provide several health care services [94,95,161-165].

The use of GenAI in some instances of service automation and guidance may be in its infancy but is encouraging. Providers are trying to explore unique ways to use AI, which requires a set of steps such as understanding the current workflow and the changes needed or aspirational workflows and aligning or designing GenAI to help in the workflow. This is similar to modifying restaurant food delivery options to suit drive-in rather than sit-in options. The providers need some work to fully automate, streamline, or re-engineer the service functions using GenAI in the future.

Summary of Findings

To summarize our findings, in this study, we conducted a systematic scoping review of the literature on how GenAI is used in clinical settings by synthesizing evidence on its application to assist, guide, and automate clinical activities and functions. Of the 161 articles, 141 (87.6%) reported using GenAI to assist services through knowledge access, collation, and filtering. The assistance of GenAI was used for disease detection (19/161, 11.8%), diagnosis (14/161, 8.7%), and screening processes (12/161, 7.5%) in the areas of radiology (17/161, 10.6%), cardiology (12/161, 7.5%), gastrointestinal medicine (4/161, 2.5%), and diabetes (6/161, 3.7%). Thus, we conclude that GenAI mainly informs rather than assisting and automating service functions. Presumably, the potential in clinical service is there, but it has yet to be actualized for GenAI.

Robustness Check Using Additional Database Search

To ensure the comprehensiveness and robustness of our findings, we expanded the search to Web of Science using similar keywords and strategies (suggested by the review team). We used the same keyword, "artificial intelligence," in all text fields over the sampling period between January 1, 2020, and November 27, 2023. Our search was restricted to peer-reviewed academic journal articles written in English. We used the Web of Science–provided "Highly Cited Papers" criterion as a

filtering mechanism to follow influential papers. Given the nonclinical context of the journals in the database, we believe that filtering based on the article's importance is reasonable. Initial search results included 1958 articles from the Web of Science Core Collection. The preliminary analysis of the annual breakdown comprised 414 articles in 2023, a total of 651 articles in 2022, a total of 519 articles in 2021, and a total of 374 articles in 2020. The search results were further reduced by removing PubMed articles for redundancy, resulting in 1221 articles.

Next, Web of Science journals include medical, nonmedical, and other clinical journals. Thus, we used simple keywords for filtering nonmedical and clinical contexts. We used the keywords "medical" and "health" mentioned in the abstract, which led to 133 articles. Finally, we read the abstracts and titles to exclude survey or meta-review and nonclinical studies. This process further narrowed down the selection to 51 relevant articles. Using ChatGPT-3.5 on November 27, 2023, we applied one-shot learning by providing 3 class definitions. We asked ChatGPT-3.5 to classify the article's abstract, with 63% (32/51) in the *assist* category, 29% (15/51) in the *guide* category, and 8% (4/51) in the *automated* category. Diagnostic assistance articles dominated, similar to the results from PubMed. However, the other categories—prescriptive guidance and clinical service recommendations—were slightly higher. This difference is explained by the nonmedical and clinical nature of the journals included in the database. The "applied" nature of the journals is more likely to explore prescriptive guidance and clinical service recommendation use cases.

Discussion

Principal Findings

This study asked RQs about how GenAI is used, with evidence, to shape health care services. It showed that 11.8% (19/161) of the studies were on automation and guidance, whereas 87.6% (141/161) reflected the assistance role of GenAI. These findings are essential to discuss and distinguish between the optimism and actual use of GenAI in health care.

Study Implications

The aspiration that GenAI has the potential to change health care significantly needs a careful revisit. Health care organizations need to assess the actual ground use for GenAI and prepare for and understand the exciting possibilities with a cautious approach rather than overly high expectations. Concerns related to the cost, privacy, misuse, and regulatory aspects of implementing and using GenAI [24-26] will become more pronounced, particularly when there is a perceived overreliance without clear promising results or actual practical use [26].

The literature synthesis in this study suggests that GenAI is mainly used for screening and diagnostic purposes using knowledge access; diagnostic processes such as predicted disease outcomes, survival, or disease classification; and improvement of the accuracy of diagnosis. This solves the problem of knowledge being available and accessible in time in a well-articulated manner to provide or render the services. This could help health care providers make more accurate and timely diagnoses, leading to earlier treatment and better patient outcomes. Such knowledge distillation helps improve diagnostic accuracy through GenAI, which can provide enough knowledge to physicians during service encounters; however, this is not hugely oriented toward higher value creation in health care.

The research synthesis also suggests that there has been some use of GenAI during different steps and aspects of guiding the service delivery processes. Still, such use could be more encouraging and significant across the board. Plausibly, GenAI can analyze large amounts of disparate data from patients to suggest personalized medicine—which may help inform treatment plans for individuals. Service delivery needs some guidance or step-by-step help to be efficient and meet the duration or time requirements to render the clinical service on time, which GenAI may solve. However, we have not yet found strong evidence for such use by any health system.

Currently, the automation of service functions using GenAI has only seen minimal instances and is yet to see widespread implementation. Automation helps offset some manual activities. However, automation may help in service functions' cost, efficiency, and flexibility while maintaining some standards across similar services.

Similarly, although we did not consider this area in the synthesis as it was out of the scope of services, GenAI can also be used in drug development and clinical trial pathways—a value proposition yet to be seen in practice. However, we do not undermine that many laboratories and pharmaceutical companies have used machine learning and AI tools and techniques in drug development and clinical trials. However, reported commercial GenAI use has not come to the limelight.

Some other plausible uses of GenAI in health care include managing supply chain data, managing medical equipment assets, maintaining gadgets and equipment, and building a robust intelligent information infrastructure to support several other activities. For example, active efforts are being undertaken to incorporate GenAI, especially in administrative use cases such as the In Basket patient messaging applications. However, assessing the clinical accuracy of such tools remains a concern.

In addition, we must incorporate user-centered design and sociotechnical frameworks into designing and building GenAI for health care use cases, for instance, to explore how GenAI can prevent a common pitfall of developing models opportunistically—based on data availability or end-point labels, adopting a user-centered design framework is vital for GenAI tools [171]. Similarly, scientific or research-oriented use of GenAI for knowledge search, articulation, or synthesis is helpful [172]. However, how far that will translate to the transformative clinical health care delivery processes while creating higher-order organizational capabilities to create value remains a concern [173].

Limitations of the Study and Scope for Future Research

Several limitations and constraints affect the interpretation and generalizability of the findings of this study. Some of these limitations indicate the need for future research in relevant areas that we discuss further. First, the study's findings were

constrained by the availability of relevant and high-quality publications and the exclusion of preprints and unpublished data to limit the specifically designed scope of the study on using GenAI in health care clinical services, which influences the comprehensiveness and accuracy of the review. There also might be a tendency for studies with positive or significant results to be published, leading to a potential publication bias. In addition, harmful or neutral findings may not be adequately represented in the review, influencing the overall assessment of GenAI's effectiveness in health care. Research should focus on patient-centered outcomes, including patient satisfaction and engagement and the impact of GenAI on the patient-provider relationship. Understanding the patient perspective is crucial for successfully integrating AI technologies into health care.

Second, the field of GenAI in health care is rapidly advancing, and new technologies and applications are continuously emerging. The findings of this study might not capture the most recent developments, and the ’conclusions of this study may become outdated quickly, specifically when some technologies have the potential to be adopted beyond institutional mechanisms, such as using GenAI mobile apps to scan images for retinopathy. Furthermore, an in-depth analysis of specific GenAI applications may open newer directions, and future research should focus on specific GenAI applications to provide detailed insights into their effectiveness and limitations. This could include applications such as diagnostic tools, treatment planning algorithms, and predictive analytics. Such heterogeneity of GenAI in health care encompasses a wide range of applications, and investigating these could make it challenging to draw overarching conclusions about GenAI’s impact on clinical services.

Third, this review may not comprehensively address ethical considerations and potential biases in the use of GenAI in health care. Ethical issues related to data privacy, algorithmic bias, and the responsible deployment of AI technologies may require more in-depth exploration. Future research should systematically explore the ethical considerations associated with GenAI use in health care. This includes issues related to data privacy, consent, transparency, and the ethical deployment of AI algorithms in clinical settings. Finally, more data, papers, articles, and longitudinal developments on some applications may enrich this study and enhance its current limited generalizability. Longitudinal studies are needed to track the impact of GenAI in health care over an extended period. This will help researchers understand the sustained effects, identify potential challenges that may arise over time, and assess the scalability and adaptability of these technologies.

Future studies could undertake comparative effectiveness research to assess how GenAI compares with traditional approaches in health care. Understanding the relative advantages and disadvantages will contribute to evidence-based decision-making. In addition, it is not clear what and how to measure the GenAI applications’ effectiveness in clinical services, leading to a call for standardized study metrics that can incorporate outcome measures and evaluation frameworks. Future research should investigate how the integration of GenAI into clinical health care services affects the workflow of health care providers. This includes understanding the time savings, challenges, and potential improvements in decision-making processes. By addressing these areas, future research can contribute to a more comprehensive understanding of the role, challenges, and potential benefits of GenAI in clinical health care services.

Actionable Policy and Practice Recommendations

The proliferation of technology often outpaces the development of appropriate regulatory and policy frameworks that are necessary for guiding proper dissemination. Our call is that, given that GenAI is emerging, policy agencies and health care organizations play a role in proactively guiding the use of GenAI in health care organizations.

What are some actionable steps for stakeholders, including health care organizations and policy makers, to navigate the integration of GenAI in health care? For health care organizations, the steps may include conducting a technology assessment vis-à-vis goals to achieve outcomes from GenAI. Evaluating the existing infrastructure and technological capabilities within the health care organization to determine readiness for GenAI integration is a first step. This will provide an understanding of the current state of technology and ensure that the necessary upgrades or modifications can be implemented to support GenAI applications, thus garnering the benefits of GenAI.

The second step is to invest in staff training and education through the development of training programs to enhance the skills of health care professionals in understanding and using GenAI technologies. Well-trained staff is essential for the effective and ethical implementation of GenAI, fostering a culture of continuous learning and adaptability. Third, health care organizations need to develop and communicate clear protocols and guidelines for the use of GenAI in different health care services, outlining ethical considerations, data privacy measures, and accountability standards. Transparent protocols help ensure the responsible and standardized use of GenAI, fostering trust among health care professionals and patients.

Fourth, health care organizations need to engage in research on GenAI through collaboration with research institutions and industry partners to participate actively in studies evaluating the effectiveness and impact of GenAI applications in specific health care domains. Involvement in research contributes to the evidence base, informs best practices, and positions the organization as a leader in health care innovation. Finally, as mentioned previously, implementing the gradual integration of GenAI rather than jumping into irrational decisions is a caution. All health systems need to gradually plan and introduce GenAI technologies, starting with pilot programs in specific departments or use cases. Gradual integration allows for careful monitoring of performance, identification of potential challenges, and iterative improvement before broader implementation.

For policy makers, much work must be done at the regulatory framework level to realize GenAI better. Policy makers must establish clear and adaptive regulatory frameworks that address the unique challenges GenAI poses in health care, ensuring patient safety, data privacy, and ethical use. There is a concern

that bias in GenAI algorithms could lead to discrimination in care delivery across patients, and the role of policy guidelines in this aspect to train and use GenAI appropriately is critical. Policy frameworks must be developed to ensure less risk, safe and ethical use, and responsible effectiveness of GenAI. Policy and industry partnerships among experts to determine relevant frameworks are vital to guide the future of GenAI to help transform health care. Robust regulations will provide a foundation for the responsible and standardized integration of GenAI technologies. An underlying challenge of GenAI is integrating it across different legacy IT systems, which involves developing and adopting interoperability standards to ensure seamless communication and data exchange between different GenAI applications and existing health care systems. Interoperability enhances efficiency, reduces redundancy, and facilitates the integration of diverse GenAI solutions. In this process, creating incentives for responsible innovation for ethical considerations and the continuous improvement of GenAI applications will drive a culture of responsibility and quality improvement, aligning technological advancements with societal needs.

Policy-level efforts also need to be oriented to allocate resources to enhance health care infrastructure, including robust connectivity and data storage capabilities, to support the data-intensive nature of GenAI applications. Adequate infrastructure is crucial for the reliable and secure functioning of GenAI in health care. Many of these enhancements may require collaboration between public health care systems, private organizations, and academia to leverage collective expertise and resources for GenAI research, development, and implementation. Finally, policies that address potential biases in GenAI applications and ensure equitable access to these technologies across diverse populations are necessary to help with proactive measures to prevent the exacerbation of existing health care disparities through the adoption of GenAI.

Conclusions

GenAI is both a tool and a complex technology. Complexity is the basis for GenAI, and thus, the use of GenAI in health care creates a set of unparalleled challenges. GenAI is costly to implement and integrate across all aspects of a health system [174]. In envisioning the future of GenAI in health care, we glimpse a transformative landscape in which technology and compassion converge for the betterment of humanity. As we stand at the intersection of innovation and responsibility, the prospect of GenAI holds immense promise in revolutionizing health care, shaping a future in which personalized, efficient, and equitable clinical services are not just aspirations but tangible realities. Our vision embraces a symbiotic relationship between technology and human touch, recognizing that the power of GenAI lies not only in its computational prowess but also in its potential to amplify the capabilities of health care professionals. Picture a world in which diagnostic accuracy is elevated, treatment plans are truly personalized, and each patient's journey is marked by precision and empathy.

Crucially, this vision hinges on responsible adoption. We envisage a future in which regulatory frameworks ensure the ethical use of GenAI, safeguard patient privacy, and uphold the principles of equity. It is a future in which interdisciplinary collaboration flourishes, bridging the expertise of health care providers, policy makers, technologists, and ethicists to navigate the complexities of this evolving landscape.

In the future, the impact of AI on human lives will be profound. Patients experience a health care system that not only heals but also understands, a system in which the integration of GenAI contributes to quicker diagnoses, more effective treatments, and improved outcomes. The human experience is at the forefront—GenAI becomes a tool for health care professionals to better connect with patients and spend more time understanding their unique needs, fears, and hopes. As we embark on this journey, it is crucial to remember that the heart of health care lies in the compassion, empathy, and wisdom of its human stewards. GenAI catalyzes empowerment, freeing health care professionals from mundane tasks to engage in meaningful interactions. It fosters a health care culture in which technology serves humanity, and the collective mission is to enhance the quality of care and life.

In embracing this vision, we are not just architects of technological progress but also custodians of a future in which GenAI and human touch coalesce to redefine health care possibilities. Let our strides be guided by a commitment to responsible innovation, a dedication to inclusivity, and an unwavering focus on the well-being of those we serve. The future of GenAI in health care is not just a scientific evolution, but it is a narrative of healing; compassion; and a shared commitment to a healthier, more humane world. However, without enough evidence, we are skeptical about the current euphoria regarding GenAI in health care.

This systematic narrative review of the preliminary evidence of using GenAI in health care clinical services provides valuable insights into the evolving landscape of AI applications in health care. The existing literature synthesis reveals promising advancements and critical considerations for integrating GenAI into clinical settings. The positive evidence underscores the potential of GenAI to revolutionize health care by offering personalized treatment plans, enhancing diagnostic accuracy, and contributing to the development of innovative therapeutic solutions. The applications of GenAI in areas such as pathology assistance, oncology decision support, and medical imaging interpretation showcase its capacity to augment health care professionals' capabilities and improve patient outcomes.

However, this review also highlights several limitations and challenges that warrant careful consideration. Issues such as the quality of available data, the rapid pace of technological evolution, and the potential for algorithmic bias highlight the complexities associated with adopting GenAI in health care. Ethical concerns, data privacy considerations, and the need for transparent guidelines underscore the importance of a thoughtful and measured approach to integration.

As we navigate the preliminary evidence, it becomes evident that a collaborative effort is required among health care organizations, policy makers, researchers, and technology developers. Establishing clear regulatory frameworks, fostering interdisciplinary collaboration, and prioritizing ethical considerations are crucial steps in ensuring the responsible

deployment of GenAI. Addressing the identified limitations through targeted research initiatives, ongoing evaluation, and continuous improvement will be essential for maximizing the benefits of GenAI while mitigating potential risks.

Moving forward, it is imperative to recognize that integrating GenAI into health care is dynamic and evolving. Future research should focus on refining our understanding of the long-term impact, patient-centered outcomes, and scalability of GenAI applications. By collectively addressing the challenges outlined in this review, stakeholders can contribute to a health care landscape in which GenAI is a powerful ally in delivering personalized, efficient, and equitable clinical services.

Acknowledgments

JK expressly acknowledges the Health Administration Research Consortium at the Business School of the University of Colorado Denver for providing a platform for the stimulating discussion and insights on this topic. The authors acknowledge Mr Bhanukesh Balabhadrapatruni, graduate student fellow at the Health Administration Research Consortium, for assisting with data categorization and citation listing. AM thanks the participants from the Society of Physician Entrepreneurs for their input about artificial intelligence in health care. VP thanks Dr Ron Li at Stanford Medicine for insights and a stimulating discussion on this topic. We used the generative AI tool ChatGPT (version 3.5; OpenAI) for automatic coding and checking the accuracy of the human coding process used to categorize the articles reviewed and synthesized in this study [166,167].

Conflicts of Interest

JK is an associate editor of the Journal of Medical Internet Research.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File , 31 KB - medinform_v12i1e52073_app1.docx]

Multimedia Appendix 2
Conversations with ChatGPT used in the Study.
[DOCX File , 85 KB - medinform_v12i1e52073_app2.docx]

References

1. Pasick A. Artificial intelligence glossary: neural networks and other terms explained. The New York Times. 2023. URL: https://www.nytimes.com/article/ai-artificial-intelligence-glossary.html [accessed 2024-01-29]
2. Roose K. A coming-out party for generative A.I., Silicon Valley's new craze. The New York Times. 2022 Oct. URL: https://www.nytimes.com/2022/10/21/technology/generative-ai.html [accessed 2024-01-29]
3. Karpathy A, Abeel P, Brockman G, Chen P, Cheung V, Duan Y. Generative models. Open AI. 2016. URL: https://openai.com/research/generative-models [accessed 2024-01-31]
4. Metz C. OpenAI plans to up the ante in tech's A.I. race. The New York Times. 2023. URL: https://www.nytimes.com/2023/03/14/technology/openai-gpt4-chatgpt.html#:~:text=But%20in%20the%20long%20term,Brockman%20said [accessed 2024-01-29]
5. Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. LaMDA: language models for dialog applications. arXiv Preprint posted online January 20, 2022 2020 [FREE Full text] [doi: 10.48550/arXiv.2201.08239]
6. Don't fear an ai-induced jobs apocalypse just yet: the west suffers from too little automation, not too much. The Economist. 2023. URL: https://www.economist.com/business/2023/03/06/dont-fear-an-ai-induced-jobs-apocalypse-just-yet [accessed 2024-01-29]
7. Harreis H, Koullias T, Roberts R, Te K. Generative AI: unlocking the future of fashion. McKinsey & Company. URL: https://www.mckinsey.com/industries/retail/our-insights/generative-ai-unlocking-the-future-of-fashion [accessed 2024-08-10]
8. Eapen TT, Venkataswamy L, Finkenstadt DJ, Folk J. How generative AI can augment human creativity. Harvard Business Review. 2023. URL: https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity [accessed 2024-01-29]
9. The race of the AI labs heats up: ChatGPT is not the only game in town. The Economist. 2023. URL: https://www.economist.com/business/2023/01/30/the-race-of-the-ai-labs-heats-up [accessed 2023-01-30]
10. Google Cloud brings generative AI to developers, businesses, and governments. Google. 2023. URL: https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-for-businesses-and-governments [accessed 2024-01-29]
11. Zheng D, He X, Jing J. Overview of artificial intelligence in breast cancer medical imaging. J Clin Med 2023 Jan 04;12(2):419 [FREE Full text] [doi: 10.3390/jcm12020419] [Medline: 36675348]
12. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023 Jun 27;33(6):1790-1796 [FREE Full text] [doi: 10.1007/s11695-023-06603-5] [Medline: 37106269]

13. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation 2023 Apr;185:109729. [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
14. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. Preprint posted online February 7, 2023 2023 [FREE Full text] [doi: 10.1101/2023.02.02.23285399] [Medline: 36798292]
15. Cirillo D, Núñez-Carpintero I, Valencia A. Artificial intelligence in cancer research: learning at different levels of data granularity. Mol Oncol 2021 Apr;15(4):817-829 [FREE Full text] [doi: 10.1002/1878-0261.12920] [Medline: 33533192]
16. Pedersen M, Verspoor K, Jenkinson M, Law M, Abbott DF, Jackson GD. Artificial intelligence for clinical decision support in neurology. Brain Commun 2020;2(2):fcaa096 [FREE Full text] [doi: 10.1093/braincomms/fcaa096] [Medline: 33134913]
17. Brynjolfsson E, McAfee A. Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. New York, NY: WW Norton & Company; 2014.
18. Raisch S, Krakowski S. Artificial intelligence and management: the automation–augmentation paradox. Acad Manage Rev 2021 Jan 14;46(1):192-210 [FREE Full text] [doi: 10.5465/amr.2018.0072]
19. 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] [Medline: 36988595]
20. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023 Mar 04;47(1):33 [FREE Full text] [doi: 10.1007/s10916-023-01925-4] [Medline: 36869927]
21. Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol 2022 Jan;35(1):23-32 [FREE Full text] [doi: 10.1038/s41379-021-00919-2] [Medline: 34611303]
22. Yu SH, Kim MS, Chung HS, Hwang EC, Jung SI, Kang TW, et al. Early experience with Watson for Oncology: a clinical decision-support system for prostate cancer treatment recommendations. World J Urol 2021 Feb;39(2):407-413 [FREE Full text] [doi: 10.1007/s00345-020-03214-y] [Medline: 32335733]
23. Wang Z, Keane PA, Chiang M, Cheung CY, Wong TY, Ting DS. Artificial intelligence and deep learning in ophthalmology. In: Lidströmer N, Ashrafian H, editors. Artificial Intelligence in Medicine. Cham, Switzerland: Springer; 2022:1519-1552.
24. Osinski B, BenTaieb A, Ho I, Jones RD, Joshi RP, Westley A, et al. Artificial intelligence-augmented histopathologic review using image analysis to optimize DNA yield from formalin-fixed paraffin-embedded slides. Mod Pathol 2022 Dec;35(12):1791-1803 [FREE Full text] [doi: 10.1038/s41379-022-01161-0] [Medline: 36198869]
25. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
26. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. Med Law Rev 2023 Nov 27;31(4):501-520 [FREE Full text] [doi: 10.1093/medlaw/fwad013] [Medline: 37218368]
27. Degnan AJ, Ghobadi EH, Hardy P, Krupinski E, Scali EP, Stratchko L, et al. Perceptual and interpretive error in diagnostic radiology-causes and potential solutions. Acad Radiol 2019 Jun;26(6):833-845 [FREE Full text] [doi: 10.1016/j.acra.2018.11.006] [Medline: 30559033]
28. Khanna NN, Maindarkar MA, Viswanathan V, Fernandes JF, Paul S, Bhagawati M, et al. Economics of artificial intelligence in healthcare: diagnosis vs. treatment. Healthcare (Basel) 2022 Dec 09;10(12):2493 [FREE Full text] [doi: 10.3390/healthcare10122493] [Medline: 36554017]
29. Curran JM, Meuter ML. Self-service technology adoption: comparing three technologies. J Serv Mark 2005;19(2):103-113. [doi: 10.1108/08876040510591411]
30. Choudhury V, Karahanna E. The relative advantage of electronic channels: a multidimensional view. MIS Q 2008;32(1):179. [doi: 10.2307/25148833]
31. Marzocchi GL, Zammit A. Self-scanning technologies in retail: determinants of adoption. Serv Ind J 2006 Sep;26(6):651-669 [FREE Full text] [doi: 10.1080/02642060600850790]
32. Campbell D, Frei F. Cost structure, customer profitability, and retention implications of self-service distribution channels: evidence from customer behavior in an online banking channel. Manag Sci 2010 Jan;56(1):4-24 [FREE Full text] [doi: 10.1287/mnsc.1090.1066]
33. Chen PY, Hitt LM. Measuring switching costs and the determinants of customer retention in internet-enabled businesses: a study of the online brokerage industry. Inf Syst Res 2002 Sep;13(3):255-274. [doi: 10.1287/isre.13.3.255.78]
34. Mols NP. The behavioral consequences of PC banking. Int J Bank Mark 1998;16(5):195-201 [FREE Full text] [doi: 10.1108/02652329810228190]
35. Apte UM, Vepsäläinen AP. High tech or high touch? Efficient channel strategies for delivering financial services. J Strateg Inf Syst 1993 Mar;2(1):39-54. [doi: 10.1016/0963-8687(93)90021-2]
36. Giebelhausen M, Robinson SG, Sirianni NJ, Brady MK. Touch versus tech: when technology functions as a barrier or a benefit to service encounters. J Mark 2014 Jul 01;78(4):113-124 [FREE Full text] [doi: 10.1509/jm.13.0056]
37. Selnes F, Hansen H. The potential hazard of self-service in developing customer loyalty. J Serv Res 2016 Jun 29;4(2):79-90 [FREE Full text] [doi: 10.1177/109467050142001]

38. Walker RH, Johnson LW. Why consumers use and do not use technology-enabled services. J Serv Mark 2006;20(2):125-135. [doi: 10.1108/08876040610657057]
39. Xue M, Hitt LM, Harker PT. Customer efficiency, channel usage, and firm performance in retail banking. Manuf Serv Oper Manag 2007 Oct;9(4):535-558 [FREE Full text] [doi: 10.1287/msom.1060.0135]
40. Johnson DS, Bardhi F, Dunn DT. Understanding how technology paradoxes affect customer satisfaction with self‐service technology: the role of performance ambiguity and trust in technology. Psychol Mark 2008 Apr 08;25(5):416-443 [FREE Full text] [doi: 10.1002/mar.20218]
41. Scherer A, Wünderlich NV, von Wangenheim F. The value of self-service: long-term effects of technology-based self-service usage on customer retention. MIS Q 2015 Jan 1;39(1):177-200. [doi: 10.25300/misq/2015/39.1.08]
42. Li S, Sun B, Wilcox RT. Cross-selling sequentially ordered products: an application to consumer banking services. J Mark Res 2018 Oct 10;42(2):233-239 [FREE Full text] [doi: 10.1509/jmkr.42.2.233.62288]
43. Bitner MJ, Brown SW, Meuter ML. Technology infusion in service encounters. J Acad Mark Sci 2000 Jan 01;28(1):138-149 [FREE Full text] [doi: 10.1177/0092070300281013]
44. Meuter ML, Ostrom AL, Roundtree RI, Bitner MJ. Self-service technologies: understanding customer satisfaction with technology-based service encounters. Journal of Marketing 2018 Oct 10;64(3):50-64. [doi: 10.1509/jmkg.64.3.50.18024]
45. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021:372:n160 [FREE Full text] [doi: 10.1136/bmj.n160]
46. Bitner M. Service and technology: opportunities and paradoxes. Manag Serv Qual 2001;11(6):375. [doi: 10.1108/09604520110410584]
47. Page MJ, McKenzie JA, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 2021 Apr;88:105906 [FREE Full text] [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
48. Baker DW. Introducing CiteScore, our journal's preferred citation index: moving beyond the impact factor. Jt Comm J Qual Patient Saf 2020 Jun;46(6):309-310 [FREE Full text] [doi: 10.1016/j.jcjq.2020.03.005] [Medline: 32402761]
49. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv 2020 Jun 12;53(3):1-34 [FREE Full text] [doi: 10.1145/3386252]
50. Dong D, Fang MJ, Tang L, Shan XH, Gao JB, Giganti F, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. Ann Oncol 2020 Jul;31(7):912-920 [FREE Full text] [doi: 10.1016/j.annonc.2020.04.003] [Medline: 32304748]
51. Ruscitti P, Bruno F, Berardicurti O, Acanfora C, Pavlych V, Palumbo P, et al. Lung involvement in macrophage activation syndrome and severe COVID-19: results from a cross-sectional study to assess clinical, laboratory and artificial intelligence-radiological differences. Ann Rheum Dis 2020 Sep;79(9):1152-1155 [FREE Full text] [doi: 10.1136/annrheumdis-2020-218048] [Medline: 32719039]
52. Shao L, Yan Y, Liu Z, Ye X, Xia H, Zhu X, et al. Radiologist-like artificial intelligence for grade group prediction of radical prostatectomy for reducing upgrading and downgrading from biopsy. Theranostics 2020;10(22):10200-10212 [FREE Full text] [doi: 10.7150/thno.48706] [Medline: 32929343]
53. Liu X, Zhang D, Liu Z, Li Z, Xie P, Sun K, et al. Deep learning radiomics-based prediction of distant metastasis in patients with locally advanced rectal cancer after neoadjuvant chemoradiotherapy: a multicentre study. EBioMedicine 2021 Jul;69:103442 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103442] [Medline: 34157487]
54. Gitto S, Cuocolo R, Annovazzi A, Anelli V, Acquasanta M, Cincotta A, et al. CT radiomics-based machine learning classification of atypical cartilaginous tumours and appendicular chondrosarcomas. EBioMedicine 2021 Jun;68:103407 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103407] [Medline: 34051442]
55. Zhang J, Yao K, Liu P, Liu Z, Han T, Zhao Z, et al. A radiomics model for preoperative prediction of brain invasion in meningioma non-invasively based on MRI: a multicentre study. EBioMedicine 2020 Aug;58:102933 [FREE Full text] [doi: 10.1016/j.ebiom.2020.102933] [Medline: 32739863]
56. Hindocha S, Charlton TG, Linton-Reid K, Hunter B, Chan C, Ahmed M, et al. A comparison of machine learning methods for predicting recurrence and death after curative-intent radiotherapy for non-small cell lung cancer: development and validation of multivariable clinical prediction models. EBioMedicine 2022 Mar;77:103911 [FREE Full text] [doi: 10.1016/j.ebiom.2022.103911] [Medline: 35248997]
57. Feng L, Liu Z, Li C, Li Z, Lou X, Shao L, et al. Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: a multicentre observational study. Lancet Digit Health 2022 Jan;4(1):e8-17 [FREE Full text] [doi: 10.1016/S2589-7500(21)00215-6] [Medline: 34952679]
58. Seah JC, Tang CH, Buchlak QD, Holt XG, Wardman JB, Aimoldin A, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health 2021 Aug;3(8):e496-e506 [FREE Full text] [doi: 10.1016/S2589-7500(21)00106-0] [Medline: 34219054]
59. Fontanellaz M, Ebner L, Huber A, Peters A, Löbelenz L, Hourscht C, et al. A deep-learning diagnostic support system for the detection of COVID-19 using chest radiographs: a multireader validation study. Invest Radiol 2021 Jun 01;56(6):348-356 [FREE Full text] [doi: 10.1097/RLI.0000000000000748] [Medline: 33259441]

60. Gu J, Tong T, Xu D, Cheng F, Fang C, He C, et al. Deep learning radiomics of ultrasonography for comprehensively predicting tumor and axillary lymph node status after neoadjuvant chemotherapy in breast cancer patients: A multicenter study. Cancer 2023 Feb 01;129(3):356-366. [doi: 10.1002/cncr.34540] [Medline: 36401611]
61. Jiang M, Li CL, Luo XM, Chuan ZR, Lv WZ, Li X, et al. Ultrasound-based deep learning radiomics in the assessment of pathological complete response to neoadjuvant chemotherapy in locally advanced breast cancer. Eur J Cancer 2021 Apr;147:95-105. [doi: 10.1016/j.ejca.2021.01.028] [Medline: 33639324]
62. Zhang Y, Liu M, Zhang L, Wang L, Zhao K, Hu S, et al. Comparison of chest radiograph captions based on natural language processing vs completed by radiologists. JAMA Netw Open 2023 Feb 01;6(2):e2255113 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.55113] [Medline: 36753278]
63. Yoon AP, Lee YL, Kane RL, Kuo C, Lin C, Chung KC. Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. JAMA Netw Open 2021 May 03;4(5):e216096 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.6096] [Medline: 33956133]
64. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. JAMA Netw Open 2020 Sep 01;3(9):e2017135 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.17135] [Medline: 32970157]
65. Zhong L, Dong D, Fang X, Zhang F, Zhang N, Zhang L, et al. A deep learning-based radiomic nomogram for prognosis and treatment decision in advanced nasopharyngeal carcinoma: a multicentre study. EBioMedicine 2021 Aug;70:103522 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103522] [Medline: 34391094]
66. Lu MT, Raghu VK, Mayrhofer T, Aerts HJ, Hoffmann U. Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. Ann Intern Med 2020 Nov 03;173(9):704-713 [FREE Full text] [doi: 10.7326/M20-1868] [Medline: 32866413]
67. Ahn JS, Ebrahimian S, McDermott S, Lee S, Naccarato L, Di Capua JF, et al. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. JAMA Netw Open 2022 Aug 01;5(8):e2229289 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.29289] [Medline: 36044215]
68. Upton R, Mumith A, Beqiri A, Parker A, Hawkes W, Gao S, et al. Automated echocardiographic detection of severe coronary artery disease using artificial intelligence. JACC Cardiovasc Imaging 2022 May;15(5):715-727 [FREE Full text] [doi: 10.1016/j.jcmg.2021.10.013] [Medline: 34922865]
69. Kusunose K, Abe T, Haga A, Fukuda D, Yamada H, Harada M, et al. A deep learning approach for assessment of regional wall motion abnormality from echocardiographic images. JACC Cardiovasc Imaging 2020 Feb;13(2 Pt 1):374-381 [FREE Full text] [doi: 10.1016/j.jcmg.2019.02.024] [Medline: 31103590]
70. Ko WY, Siontis KC, Attia ZI, Carter RE, Kapa S, Ommen SR, et al. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. J Am Coll Cardiol 2020 Feb 25;75(7):722-733 [FREE Full text] [doi: 10.1016/j.jacc.2019.12.030] [Medline: 32081280]
71. Vaid A, Johnson KW, Badgeley MA, Somani SS, Bicak M, Landi I, et al. Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. JACC Cardiovasc Imaging 2022 Mar;15(3):395-410 [FREE Full text] [doi: 10.1016/j.jcmg.2021.08.004] [Medline: 34656465]
72. Elias P, Poterucha TJ, Rajaram V, Moller LM, Rodriguez V, Bhave S, et al. Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease. J Am Coll Cardiol 2022 Aug 09;80(6):613-626 [FREE Full text] [doi: 10.1016/j.jacc.2022.05.029] [Medline: 35926935]
73. Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med 2021 May;27(5):815-819. [doi: 10.1038/s41591-021-01335-4] [Medline: 33958795]
74. Wu S, Chen X, Pan J, Dong W, Diao X, Zhang R, et al. An artificial intelligence system for the detection of bladder cancer via cystoscopy: a multicenter diagnostic study. J Natl Cancer Inst 2022 Feb 07;114(2):220-227 [FREE Full text] [doi: 10.1093/jnci/djab179] [Medline: 34473310]
75. Narang A, Bae R, Hong H, Thomas Y, Surette S, Cadieu C, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. JAMA Cardiol 2021 Jun 01;6(6):624-632 [FREE Full text] [doi: 10.1001/jamacardio.2021.0185] [Medline: 33599681]
76. Yuan XL, Guo LJ, Liu W, Zeng X, Mou Y, Bai S, et al. Artificial intelligence for detecting superficial esophageal squamous cell carcinoma under multiple endoscopic imaging modalities: a multicenter study. J Gastroenterol Hepatol 2022 Jan;37(1):169-178 [FREE Full text] [doi: 10.1111/jgh.15689] [Medline: 34532890]
77. Attia ZI, Kapa S, Dugan J, Pereira N, Noseworthy PA, Jimenez FL, Discover Consortium (DigitalNoninvasive Screening for COVID-19 with AI ECG Repository). Rapid exclusion of COVID infection with the artificial intelligence electrocardiogram. Mayo Clin Proc 2021 Aug;96(8):2081-2094 [FREE Full text] [doi: 10.1016/j.mayocp.2021.05.027] [Medline: 34353468]
78. Kashou AH, Medina-Inojosa JR, Noseworthy PA, Rodeheffer RJ, Lopez-Jimenez F, Attia IZ, et al. Artificial intelligence-augmented electrocardiogram detection of left ventricular systolic dysfunction in the general population. Mayo Clin Proc 2021 Oct;96(10):2576-2586 [FREE Full text] [doi: 10.1016/j.mayocp.2021.02.029] [Medline: 34120755]

79. Kwon JM, Kim KH, Medina-Inojosa J, Jeon KH, Park J, Oh BH. Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography. J Heart Lung Transplant 2020 Aug;39(8):805-814. [doi: 10.1016/j.healun.2020.04.009] [Medline: 32381339]
80. Asch FM, Mor-Avi V, Rubenson D, Goldstein S, Saric M, Mikati I, et al. Deep learning-based automated echocardiographic quantification of left ventricular ejection fraction: a point-of-care solution. Circ Cardiovasc Imaging 2021 Jun;14(6):e012293. [doi: 10.1161/CIRCIMAGING.120.012293] [Medline: 34126754]
81. Kashou AH, Rabinstein AA, Attia IZ, Asirvatham SJ, Gersh BJ, Friedman PA, et al. Recurrent cryptogenic stroke: a potential role for an artificial intelligence-enabled electrocardiogram? HeartRhythm Case Rep 2020 Apr;6(4):202-205 [FREE Full text] [doi: 10.1016/j.hrcr.2019.12.013] [Medline: 32322497]
82. Wu L, He X, Liu M, Xie H, An P, Zhang J, et al. Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial. Endoscopy 2021 Dec;53(12):1199-1207. [doi: 10.1055/a-1350-5583] [Medline: 33429441]
83. Yang X, Wang H, Dong Q, Xu Y, Liu H, Ma X, et al. An artificial intelligence system for distinguishing between gastrointestinal stromal tumors and leiomyomas using endoscopic ultrasonography. Endoscopy 2022 Mar;54(3):251-261. [doi: 10.1055/a-1476-8931] [Medline: 33827140]
84. Herrin J, Abraham NS, Yao X, Noseworthy PA, Inselman J, Shah ND, et al. Comparative effectiveness of machine learning approaches for predicting gastrointestinal bleeds in patients receiving antithrombotic treatment. JAMA Netw Open 2021 May 03;4(5):e2110703 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10703] [Medline: 34019087]
85. Xie X, Xiao YF, Zhao XY, Li JJ, Yang QQ, Peng X, et al. Development and validation of an artificial intelligence model for small bowel capsule endoscopy video review. JAMA Netw Open 2022 Jul 01;5(7):e2221992 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.21992] [Medline: 35834249]
86. Shung DL, Au B, Taylor RA, Tay JK, Laursen SB, Stanley AJ, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. Gastroenterology 2020 Jan;158(1):160-167 [FREE Full text] [doi: 10.1053/j.gastro.2019.09.009] [Medline: 31562847]
87. Bhuiyan A, Govindaiah A, Deobhakta A, Gupta M, Rosen R, Saleem S, et al. Development and validation of an automated diabetic retinopathy screening tool for primary care setting. Diabetes Care 2020 Oct;43(10):e147-e148 [FREE Full text] [doi: 10.2337/dc19-2133] [Medline: 32855159]
88. Heydon P, Egan C, Bolter L, Chambers R, Anderson J, Aldington S, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. Br J Ophthalmol 2021 May;105(5):723-728 [FREE Full text] [doi: 10.1136/bjophthalmol-2020-316594] [Medline: 32606081]
89. Olvera-Barrios A, Heeren TF, Balaskas K, Chambers R, Bolter L, Egan C, et al. Diagnostic accuracy of diabetic retinopathy grading by an artificial intelligence-enabled algorithm compared with a human standard for wide-field true-colour confocal scanning and standard digital retinal images. Br J Ophthalmol 2021 Feb;105(2):265-270. [doi: 10.1136/bjophthalmol-2019-315394] [Medline: 32376611]
90. Dai L, Wu L, Li H, Cai C, Wu Q, Kong H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat Commun 2021 May 28;12(1):3242 [FREE Full text] [doi: 10.1038/s41467-021-23458-5] [Medline: 34050158]
91. Ipp E, Liljenquist D, Bode B, Shah VN, Silverstein S, Regillo CD, EyeArt Study Group. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. JAMA Netw Open 2021 Nov 01;4(11):e2134254 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.34254] [Medline: 34779843]
92. Ravaut M, Harish V, Sadeghi H, Leung KK, Volkovs M, Kornas K, et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. JAMA Netw Open 2021 May 03;4(5):e2111315 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11315] [Medline: 34032855]
93. Bachar N, Benbassat D, Brailovsky D, Eshel Y, Glück D, Levner D, et al. An artificial intelligence-assisted diagnostic platform for rapid near-patient hematology. Am J Hematol 2021 Oct 01;96(10):1264-1274 [FREE Full text] [doi: 10.1002/ajh.26295] [Medline: 34264525]
94. Dong L, He W, Zhang R, Ge Z, Wang YX, Zhou J, et al. Artificial intelligence for screening of multiple retinal and optic nerve diseases. JAMA Netw Open 2022 May 02;5(5):e229960 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.9960] [Medline: 35503220]
95. Lee AY, Yanagihara RT, Lee CS, Blazes M, Jung HC, Chee YE, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care 2021 May;44(5):1168-1175 [FREE Full text] [doi: 10.2337/dc20-1877] [Medline: 33402366]
96. Lee Y, Kim G, Jun JE, Park H, Lee WJ, Hwang YC, et al. An integrated digital health care platform for diabetes management with ai-based dietary management: 48-week results from a randomized controlled trial. Diabetes Care 2023 May 01;46(5):959-966. [doi: 10.2337/dc22-1929] [Medline: 36821833]
97. Oikonomidi T, Ravaud P, Cosson E, Montori V, Tran VT. Evaluation of patient willingness to adopt remote digital monitoring for diabetes management. JAMA Netw Open 2021 Jan 04;4(1):e2033115 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.33115] [Medline: 33439263]

98. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020 Aug;159(2):512-20.e7. [doi: 10.1053/j.gastro.2020.04.062] [Medline: 32371116]
99. Wang P, Liu P, Glissen Brown JR, Berzin TM, Zhou G, Lei S, et al. Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs routine white-light colonoscopy in a prospective tandem study. Gastroenterology 2020 Oct;159(4):1252-61.e5. [doi: 10.1053/j.gastro.2020.06.023] [Medline: 32562721]
100. Svoboda E. Artificial intelligence is improving the detection of lung cancer. Nature 2020 Nov;587(7834):S20-S22. [doi: 10.1038/d41586-020-03157-9] [Medline: 33208974]
101. Song Z, Zou S, Zhou W, Huang Y, Shao L, Yuan J, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 2020 Aug 27;11(1):4294 [FREE Full text] [doi: 10.1038/s41467-020-18147-8] [Medline: 32855423]
102. Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel AS, Naheyan T, et al. Tuberculosis detection from chest x-rays for triaging in a high tuberculosis-burden setting: an evaluation of five artificial intelligence algorithms. Lancet Digit Health 2021 Sep;3(9):e543-e554 [FREE Full text] [doi: 10.1016/S2589-7500(21)00116-3] [Medline: 34446265]
103. Tang LY, Coxson HO, Lam S, Leipsic J, Tam RC, Sin DD. Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT. Lancet Digit Health 2020 May;2(5):e259-e267 [FREE Full text] [doi: 10.1016/S2589-7500(20)30064-9] [Medline: 33328058]
104. Kim H, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Health 2020 Mar;2(3):e138-e148. [doi: 10.1016/s2589-7500(20)30003-0]
105. Wu S, Hong G, Xu A, Zeng H, Chen X, Wang Y, et al. Artificial intelligence-based model for lymph node metastases detection on whole slide images in bladder cancer: a retrospective, multicentre, diagnostic study. Lancet Oncol 2023 Apr;24(4):360-370. [doi: 10.1016/S1470-2045(23)00061-X] [Medline: 36893772]
106. Weigt J, Repici A, Antonelli G, Afifi A, Kliegis L, Correale L, et al. Performance of a new integrated computer-assisted system (CADe/CADx) for detection and characterization of colorectal neoplasia. Endoscopy 2022 Feb;54(2):180-184. [doi: 10.1055/a-1372-0419] [Medline: 33494106]
107. Homayounieh F, Digumarthy S, Ebrahimian S, Rueckel J, Hoppe BF, Sabel BO, et al. An artificial intelligence-based chest X-ray model on human nodule detection accuracy from a multicenter study. JAMA Netw Open 2021 Dec 01;4(12):e2141096 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.41096] [Medline: 34964851]
108. Glissen Brown JR, Mansour NM, Wang P, Chuchuca MA, Minchenberg SB, Chandnani M, et al. Deep learning computer-aided polyp detection reduces adenoma miss rate: a united states multi-center randomized tandem colonoscopy study (CADeT-CS Trial). Clin Gastroenterol Hepatol 2022 Jul;20(7):1499-507.e4 [FREE Full text] [doi: 10.1016/j.cgh.2021.09.009] [Medline: 34530161]
109. Foersch S, Eckstein M, Wagner DC, Gach F, Woerl AC, Geiger J, et al. Deep learning for diagnosis and survival prediction in soft tissue sarcoma. Ann Oncol 2021 Sep;32(9):1178-1187 [FREE Full text] [doi: 10.1016/j.annonc.2021.06.007] [Medline: 34139273]
110. Jin EH, Lee D, Bae JH, Kang HY, Kwak M, Seo JY, et al. Improved accuracy in optical diagnosis of colorectal polyps using convolutional neural networks with visual explanations. Gastroenterology 2020 Jun;158(8):2169-79.e8. [doi: 10.1053/j.gastro.2020.02.036] [Medline: 32119927]
111. Shi Y, Wang Z, Chen P, Cheng P, Zhao K, Zhang H, Alzheimer's Disease Neuroimaging Initiative. Episodic memory-related imaging features as valuable biomarkers for the diagnosis of Alzheimer's disease: a multicenter study based on machine learning. Biol Psychiatry Cogn Neurosci Neuroimaging 2023 Feb;8(2):171-180. [doi: 10.1016/j.bpsc.2020.12.007] [Medline: 33712376]
112. Huang B, Tian S, Zhan N, Ma J, Huang Z, Zhang C, et al. Accurate diagnosis and prognosis prediction of gastric cancer using deep learning on digital pathological images: a retrospective multicentre study. EBioMedicine 2021 Nov;73:103631 [FREE Full text] [doi: 10.1016/j.ebiom.2021.103631] [Medline: 34678610]
113. Jin C, Chen W, Cao Y, Xu Z, Tan Z, Zhang X, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 2020 Oct 09;11(1):5088 [FREE Full text] [doi: 10.1038/s41467-020-18685-1] [Medline: 33037212]
114. Goh KH, Wang L, Yeow AY, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun 2021 Jan 29;12(1):711 [FREE Full text] [doi: 10.1038/s41467-021-20910-4] [Medline: 33514699]
115. Zhou Q, Zuley M, Guo Y, Yang L, Nair B, Vargo A, et al. A machine and human reader study on AI diagnosis model safety under attacks of adversarial images. Nat Commun 2021 Dec 14;12(1):7281 [FREE Full text] [doi: 10.1038/s41467-021-27577-x] [Medline: 34907229]
116. Peng S, Liu Y, Lv W, Liu L, Zhou Q, Yang H, et al. Deep learning-based artificial intelligence model to assist thyroid nodule diagnosis and management: a multicentre diagnostic study. Lancet Digit Health 2021 Apr;3(4):e250-e259 [FREE Full text] [doi: 10.1016/S2589-7500(21)00041-8] [Medline: 33766289]

117. Pantanowitz L, Quiroga-Garza GM, Bien L, Heled R, Laifenfeld D, Linhart C, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. Lancet Digit Health 2020 Aug;2(8):e407-e416. [doi: 10.1016/s2589-7500(20)30159-x]
118. Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol 2020 Feb;21(2):222-232. [doi: 10.1016/S1470-2045(19)30738-7] [Medline: 31926806]
119. Venkatesan P. Artificial intelligence and cancer diagnosis: caution needed. Lancet Oncol 2021 Oct;22(10):1364. [doi: 10.1016/S1470-2045(21)00533-7] [Medline: 34509184]
120. Gao K, Su J, Jiang Z, Zeng L, Feng Z, Shen H, et al. Dual-branch combination network (DCN): towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Med Image Anal 2021 Jan;67:101836 [FREE Full text] [doi: 10.1016/j.media.2020.101836] [Medline: 33129141]
121. Pfob A, Sidey-Gibbons C, Barr RG, Duda V, Alwafai Z, Balleyguier C, et al. Intelligent multi-modal shear wave elastography to reduce unnecessary biopsies in breast cancer diagnosis (INSPiRED 002): a retrospective, international, multicentre analysis. Eur J Cancer 2022 Dec;177:1-14. [doi: 10.1016/j.ejca.2022.09.018] [Medline: 36283244]
122. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020 Jan;577(7788):89-94. [doi: 10.1038/s41586-019-1799-6] [Medline: 31894144]
123. Bachtiger P, Petri CF, Scott FE, Ri Park S, Kelshiker MA, Sahemey HK, et al. Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ECG-enabled stethoscope examination in London, UK: a prospective, observational, multicentre study. Lancet Digit Health 2022 Feb;4(2):e117-e125. [doi: 10.1016/s2589-7500(21)00256-9]
124. Kann BH, Likitlersuang J, Bontempi D, Ye Z, Aneja S, Bakst R, et al. Screening for extranodal extension in HPV-associated oropharyngeal carcinoma: evaluation of a CT-based deep learning algorithm in patient data from a multicentre, randomised de-escalation trial. Lancet Digit Health 2023 Jun;5(6):e360-e369 [FREE Full text] [doi: 10.1016/S2589-7500(23)00046-8] [Medline: 37087370]
125. Soltan AA, Kouchaki S, Zhu T, Kiyasseh D, Taylor T, Hussain ZB, et al. Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. Lancet Digit Health 2021 Feb;3(2):e78-e87 [FREE Full text] [doi: 10.1016/S2589-7500(20)30274-0] [Medline: 33509388]
126. Xie Y, Zhao L, Yang X, Wu X, Yang Y, Huang X, et al. Screening candidates for refractive surgery with corneal tomographic-based deep learning. JAMA Ophthalmol 2020 May 01;138(5):519-526 [FREE Full text] [doi: 10.1001/jamaophthalmol.2020.0507] [Medline: 32215587]
127. Abbasi J. Artificial intelligence improves breast cancer screening in study. JAMA 2020 Feb 11;323(6):499. [doi: 10.1001/jama.2020.0370] [Medline: 32044919]
128. Xu H, Tang RS, Lam TY, Zhao G, Lau JY, Liu Y, et al. Artificial intelligence-assisted colonoscopy for colorectal cancer screening: a multicenter randomized controlled trial. Clin Gastroenterol Hepatol 2023 Feb;21(2):337-46.e3 [FREE Full text] [doi: 10.1016/j.cgh.2022.07.006] [Medline: 35863686]
129. Sun Y, Zhang L, Dong D, Li X, Wang J, Yin C, et al. Application of an individualized nomogram in first-trimester screening for trisomy 21. Ultrasound Obstet Gynecol 2021 Jul;58(1):56-66 [FREE Full text] [doi: 10.1002/uog.22087] [Medline: 32438493]
130. Zeleznik R, Foldyna B, Eslami P, Weiss J, Alexander I, Taron J, et al. Deep convolutional neural networks to predict cardiovascular risk from computed tomography. Nat Commun 2021 Jan 29;12(1):715 [FREE Full text] [doi: 10.1038/s41467-021-20966-2] [Medline: 33514711]
131. Liu CM, Chang SL, Chen HH, Chen WS, Lin YJ, Lo LW, et al. The clinical application of the deep learning technique for predicting trigger origins in patients with paroxysmal atrial fibrillation with catheter ablation. Circ Arrhythm Electrophysiol 2020 Nov;13(11):e008518. [doi: 10.1161/CIRCEP.120.008518] [Medline: 33021404]
132. Qiang M, Li C, Sun Y, Sun Y, Ke L, Xie C, et al. A prognostic predictive system based on deep learning for locoregionally advanced nasopharyngeal carcinoma. J Natl Cancer Inst 2021 May 04;113(5):606-615 [FREE Full text] [doi: 10.1093/jnci/djaa149] [Medline: 32970812]
133. She Y, He B, Wang F, Zhong Y, Wang T, Liu Z, et al. Deep learning for predicting major pathological response to neoadjuvant chemoimmunotherapy in non-small cell lung cancer: a multicentre study. EBioMedicine 2022 Dec;86:104364 [FREE Full text] [doi: 10.1016/j.ebiom.2022.104364] [Medline: 36395737]
134. Wang L, Ding L, Liu Z, Sun L, Chen L, Jia R, et al. Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. Br J Ophthalmol 2020 Mar;104(3):318-323. [doi: 10.1136/bjophthalmol-2018-313706] [Medline: 31302629]
135. Li D, Bledsoe JR, Zeng Y, Liu W, Hu Y, Bi K, et al. A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals. Nat Commun 2020 Nov 26;11(1):6004 [FREE Full text] [doi: 10.1038/s41467-020-19817-3] [Medline: 33244018]
136. Yu G, Sun K, Xu C, Shi XH, Wu C, Xie T, et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. Nat Commun 2021 Nov 02;12(1):6311 [FREE Full text] [doi: 10.1038/s41467-021-26643-8] [Medline: 34728629]

137. Kwon JM, Cho Y, Jeon KH, Cho S, Kim KH, Baek SD, et al. A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. Lancet Digit Health 2020 Jul;2(7):e358-e367 [FREE Full text] [doi: 10.1016/S2589-7500(20)30108-4] [Medline: 33328095]
138. Lin A, Manral N, McElhinney P, Killekar A, Matsumoto H, Kwiecinski J, et al. Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study. Lancet Digit Health 2022 Apr;4(4):e256-e265 [FREE Full text] [doi: 10.1016/S2589-7500(22)00022-X] [Medline: 35337643]
139. Storelli L, Azzimonti M, Gueye M, Vizzino C, Preziosa P, Tedeschi G, et al. A deep learning approach to predicting disease progression in multiple sclerosis using magnetic resonance imaging. Invest Radiol 2022 Jul 01;57(7):423-432. [doi: 10.1097/RLI.0000000000000854] [Medline: 35093968]
140. Mao N, Zhang H, Dai Y, Li Q, Lin F, Gao J, et al. Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study. Br J Cancer 2023 Mar;128(5):793-804 [FREE Full text] [doi: 10.1038/s41416-022-02092-y] [Medline: 36522478]
141. Ueno S, Berntsen J, Ito M, Uchiyama K, Okimura T, Yabuuchi A, et al. Pregnancy prediction performance of an annotation-free embryo scoring system on the basis of deep learning after single vitrified-warmed blastocyst transfer: a single-center large cohort retrospective study. Fertil Steril 2021 Oct;116(4):1172-1180 [FREE Full text] [doi: 10.1016/j.fertnstert.2021.06.001] [Medline: 34246469]
142. Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, et al. Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study. Lancet Oncol 2021 Jan;22(1):132-141. [doi: 10.1016/S1470-2045(20)30535-0] [Medline: 33387492]
143. Li X, Gao H, Zhu J, Huang Y, Zhu Y, Huang W, et al. 3D deep learning model for the pretreatment evaluation of treatment response in esophageal carcinoma: a prospective study (ChiCTR2000039279. Int J Radiat Oncol Biol Phys 2021 Nov 15;111(4):926-935 [FREE Full text] [doi: 10.1016/j.ijrobp.2021.06.033] [Medline: 34229050]
144. Wu L, Ye W, Liu Y, Chen D, Wang Y, Cui Y, et al. An integrated deep learning model for the prediction of pathological complete response to neoadjuvant chemotherapy with serial ultrasonography in breast cancer patients: a multicentre, retrospective study. Breast Cancer Res 2022 Nov 21;24(1):81 [FREE Full text] [doi: 10.1186/s13058-022-01580-6] [Medline: 36414984]
145. Suri JS, Agarwal S, Saba L, Chabert GL, Carriero A, Paschè A, et al. Multicenter study on COVID-19 lung computed tomography segmentation with varying glass ground opacities using unseen deep learning artificial intelligence paradigms: COVLIAS 1.0 validation. J Med Syst 2022 Aug 21;46(10):62 [FREE Full text] [doi: 10.1007/s10916-022-01850-y] [Medline: 35988110]
146. Khurshid S, Friedman S, Pirruccello JP, Di Achille P, Diamant N, Anderson CD, et al. Deep learning to predict cardiac magnetic resonance-derived left ventricular mass and hypertrophy from 12-lead ECGs. Circ Cardiovasc Imaging 2021 Jun;14(6):e012281 [FREE Full text] [doi: 10.1161/CIRCIMAGING.120.012281] [Medline: 34126762]
147. Liu XP, Jin X, Seyed Ahmadian S, Yang X, Tian SF, Cai YX, et al. Clinical significance and molecular annotation of cellular morphometric subtypes in lower-grade gliomas discovered by machine learning. Neuro Oncol 2023 Jan 05;25(1):68-81 [FREE Full text] [doi: 10.1093/neuonc/noac154] [Medline: 35716369]
148. Akal F, Batu ED, Sonmez HE, Karadağ Ş, Demir F, Ayaz NA, et al. Diagnosing growing pains in children by using machine learning: a cross-sectional multicenter study. Med Biol Eng Comput 2022 Dec;60(12):3601-3614. [doi: 10.1007/s11517-022-02699-6] [Medline: 36264529]
149. Awada H, Durmaz A, Gurnari C, Kishtagari A, Meggendorfer M, Kerr CM, et al. Machine learning integrates genomic signatures for subclassification beyond primary and secondary acute myeloid leukemia. Blood 2021 Nov 11;138(19):1885-1895 [FREE Full text] [doi: 10.1182/blood.2020010603] [Medline: 34075412]
150. Moyer JD, Lee P, Bernard C, Henry L, Lang E, Cook F, Traumabase Group®. Machine learning-based prediction of emergency neurosurgery within 24 h after moderate to severe traumatic brain injury. World J Emerg Surg 2022 Aug 03;17(1):42 [FREE Full text] [doi: 10.1186/s13017-022-00449-5] [Medline: 35922831]
151. Hollon T, Jiang C, Chowdury A, Nasir-Moin M, Kondepudi A, Aabedi A, et al. Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. Nat Med 2023 Apr;29(4):828-832 [FREE Full text] [doi: 10.1038/s41591-023-02252-4] [Medline: 36959422]
152. Takenaka K, Ohtsuka K, Fujii T, Negi M, Suzuki K, Shimizu H, et al. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology 2020 Jun;158(8):2150-2157. [doi: 10.1053/j.gastro.2020.02.012] [Medline: 32060000]
153. Savage N. Why artificial intelligence needs to understand consequences. Nature (Forthcoming) 2023 Feb 24. [doi: 10.1038/d41586-023-00577-1] [Medline: 36829060]
154. -. Artificial intelligence predicts drug response. Cancer Discov 2021 Jan;11(1):4-5. [doi: 10.1158/2159-8290.CD-NB2020-109] [Medline: 33239267]
155. Wagner M, Müller-Stich BP, Kisilenko A, Tran D, Heger P, Mündermann L, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. Med Image Anal 2023 May;86:102770 [FREE Full text] [doi: 10.1016/j.media.2023.102770] [Medline: 36889206]

156. Soda P, D'Amico NC, Tessadori J, Valbusa G, Guarrasi V, Bortolotto C, et al. AIforCOVID: predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study. Med Image Anal 2021 Dec;74:102216 [FREE Full text] [doi: 10.1016/j.media.2021.102216] [Medline: 34492574]
157. Avari P, Leal Y, Herrero P, Wos M, Jugnee N, Arnoriaga-Rodríguez M, et al. Safety and feasibility of the PEPPER adaptive bolus advisor and safety system: a randomized control study. Diabetes Technol Ther 2021 Mar 01;23(3):175-186. [doi: 10.1089/dia.2020.0301] [Medline: 33048581]
158. Wathour J, Govaerts PJ, Deggouj N. From manual to artificial intelligence fitting: two cochlear implant case studies. Cochlear Implants Int 2020 Sep;21(5):299-305. [doi: 10.1080/14670100.2019.1667574] [Medline: 31530099]
159. Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. JAMA Netw Open 2021 Feb 01;4(2):e2037107 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.37107] [Medline: 33599773]
160. Eilts SK, Pfeil JM, Poschkamp B, Krohne TU, Eter N, Barth T, Comparing Alternative Ranibizumab Dosages for SafetyEfficacy in Retinopathy of Prematurity (CARE-ROP) Study Group. Assessment of retinopathy of prematurity regression and reactivation using an artificial intelligence-based vascular severity score. JAMA Netw Open 2023 Jan 03;6(1):e2251512 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.51512] [Medline: 36656578]
161. Takeda I, Yamada A, Onodera H. Artificial intelligence-assisted motion capture for medical applications: a comparative study between markerless and passive marker motion capture. Comput Methods Biomech Biomed Engin 2021 Jun;24(8):864-873. [doi: 10.1080/10255842.2020.1856372] [Medline: 33290107]
162. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, NextDREAM Consortium. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. Nat Med 2020 Sep;26(9):1380-1384. [doi: 10.1038/s41591-020-1045-7] [Medline: 32908282]
163. Carvalho DM, Richardson PJ, Olaciregui N, Stankunaite R, Lavarino C, Molinari V, et al. Repurposing Vandetanib plus everolimus for the treatment of -mutant diffuse intrinsic pontine glioma. Cancer Discov 2022 Feb;12(2):416-431 [FREE Full text] [doi: 10.1158/2159-8290.CD-20-1201] [Medline: 34551970]
164. Sheridan C. Massive data initiatives and AI provide testbed for pandemic forecasting. Nat Biotechnol 2020 Sep;38(9):1010-1013. [doi: 10.1038/s41587-020-0671-4] [Medline: 32887968]
165. Meeuws M, Pascoal D, Janssens de Varebeke S, De Ceulaer G, Govaerts PJ. Cochlear implant telemedicine: remote fitting based on psychoacoustic self-tests and artificial intelligence. Cochlear Implants Int 2020 Sep 13;21(5):260-268. [doi: 10.1080/14670100.2020.1757840] [Medline: 32397922]
166. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45-10 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
167. Dong Q, Li L, Dai D, Zheng C, Wu Z, Chang B, et al. A survey for in-context learning. arXiv. Preprint posted online December 31, 2022 2022 [FREE Full text]
168. Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, et al. Development and validation of an artificial intelligence system to optimize clinician review of patient records. JAMA Netw Open 2021 Jul 01;4(7):e2117391 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.17391] [Medline: 34297075]
169. Cowan RP, Rapoport AM, Blythe J, Rothrock J, Knievel K, Peretz AM, et al. Diagnostic accuracy of an artificial intelligence online engine in migraine: a multi-center study. Headache 2022 Jul;62(7):870-882 [FREE Full text] [doi: 10.1111/head.14324] [Medline: 35657603]
170. Curran JM, Meuter ML, Surprenant CF. Intentions to use self-service technologies: a confluence of multiple attitudes. J Serv Res 2016 Jun 29;5(3):209-224 [FREE Full text] [doi: 10.1177/1094670502238916]
171. Dabholkar PA. Consumer evaluations of new technology-based self-service options: an investigation of alternative models of service quality. Int J Res Mark 1996;13(1):29-51 [FREE Full text] [doi: 10.1016/0167-8116(95)00027-5]
172. Seneviratne MG, Li RC, Schreier M, Lopez-Martinez D, Patel BS, Yakubovich A, et al. User-centred design for machine learning in health care: a case study from care management. BMJ Health Care Inform 2022 Oct 11;29(1):e100656. [doi: 10.1136/bmjhci-2022-100656] [Medline: 36220304]
173. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. Front Digit Health 2021;3:645232 [FREE Full text] [doi: 10.3389/fdgth.2021.645232] [Medline: 34713115]
174. Novak LL, Russell RG, Garvey K, Patel M, Thomas Craig KJ, Snowdon J, et al. Clinical use of artificial intelligence requires AI-capable organizations. JAMIA Open 2023 Jul;6(2):ooad028 [FREE Full text] [doi: 10.1093/jamiaopen/ooad028] [Medline: 37152469]

Abbreviations

AI: artificial intelligence
GenAI: generative artificial intelligence tools and applications
ICL: in-context learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RQ: research question

Edited by A Castonguay; submitted 21.08.23; peer-reviewed by SH Kim, Y Wang, S Pesala; comments to author 19.09.23; revised version received 12.10.23; accepted 30.01.24; published 20.03.24.

Please cite as:
Yim D, Khuntia J, Parameswaran V, Meyers A
Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review
JMIR Med Inform 2024;12:e52073
URL: https://medinform.jmir.org/2024/1/e52073
doi:10.2196/52073
PMID:38506918

Impact of Translation on Biomedical Information Extraction: Experiment on Real-Life Clinical Notes

Christel Gérardin[1], MSc, MD; Yuhan Xiong[1,2], MS; Perceval Wajsbürt[3], PhD; Fabrice Carrat[1,4], MD, PhD; Xavier Tannier[5], PhD

1

2

3

4

5

Corresponding Author:
Christel Gérardin, MSc, MD

Abstract

Background: Biomedical natural language processing tasks are best performed with English models, and translation tools have undergone major improvements. On the other hand, building annotated biomedical data sets remains a challenge.

Objective: The aim of our study is to determine whether the use of English tools to extract and normalize French medical concepts based on translations provides comparable performance to that of French models trained on a set of annotated French clinical notes.

Methods: We compared 2 methods: 1 involving French-language models and 1 involving English-language models. For the native French method, the named entity recognition and normalization steps were performed separately. For the translated English method, after the first translation step, we compared a 2-step method and a terminology-oriented method that performs extraction and normalization at the same time. We used French, English, and bilingual annotated data sets to evaluate all stages (named entity recognition, normalization, and translation) of our algorithms.

Results: The native French method outperformed the translated English method, with an overall F_1-score of 0.51 (95% CI 0.47-0.55), compared with 0.39 (95% CI 0.34-0.44) and 0.38 (95% CI 0.36-0.40) for the 2 English methods tested.

Conclusions: Despite recent improvements in translation models, there is a significant difference in performance between the 2 approaches in favor of the native French method, which is more effective on French medical texts, even with few annotated documents.

(JMIR Med Inform 2024;12:e49607) doi:10.2196/49607

KEYWORDS

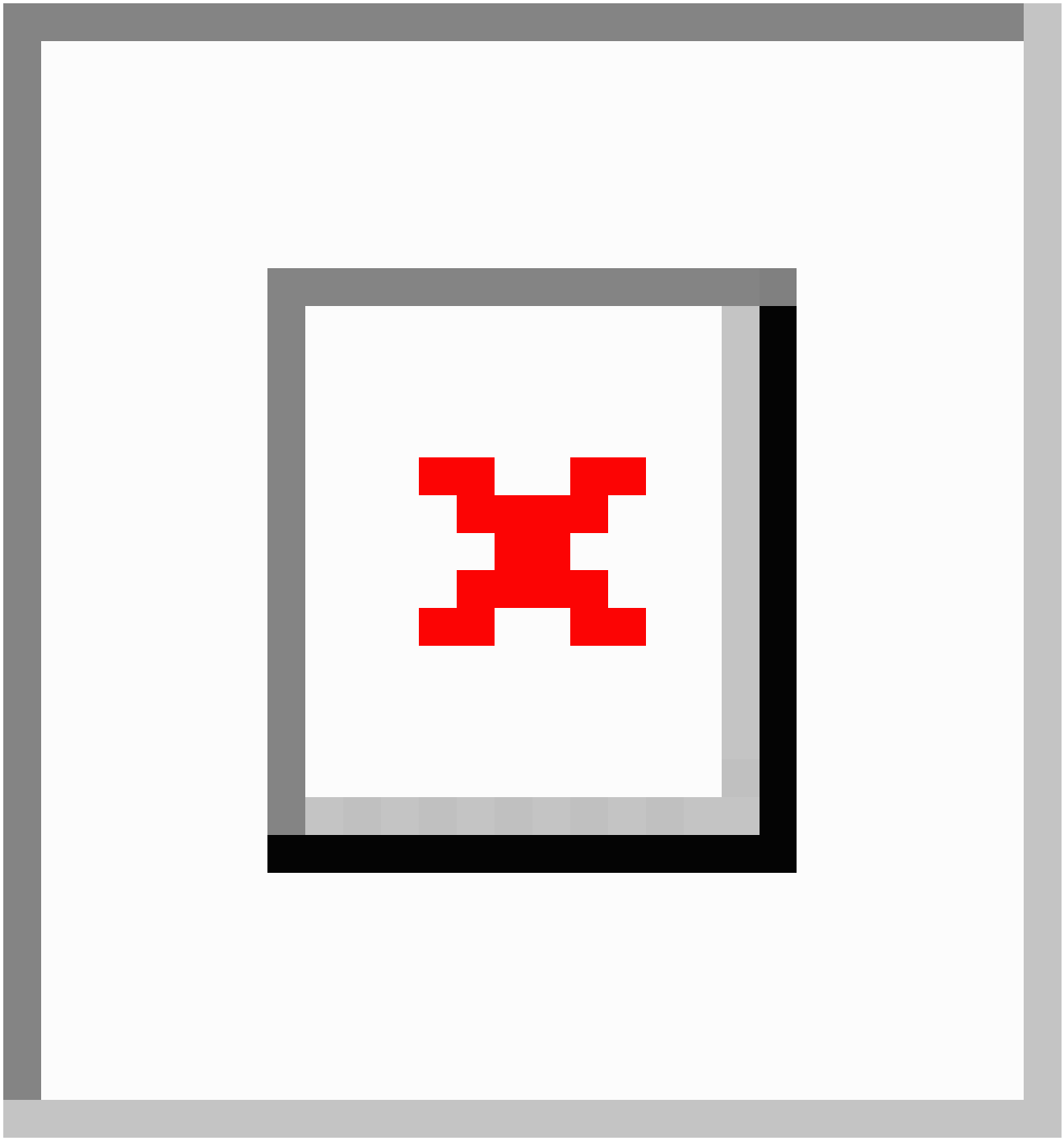
concept normalization; named entity recognition; natural language processing; translation; translational tool; biomedical data set; bilingual language model

Introduction

Named entity recognition (NER) and term normalization are important steps in biomedical natural language processing (NLP). NER is used to extract key information from textual medical reports, and normalization consists of matching a specific term to its formal reference in a shared terminology such as the United Medical Language System (UMLS) Metathesaurus [1]. Major improvements have been made recently in these areas, particularly for English, as a huge amount of data is available in the literature and resources. Modern automatic language processing relies heavily on pretrained language models, which enable efficient semantic representation of texts. The development of algorithms such as transformers [2,3] has led to significant progress in this field.

In Figure 1, the term “mention level” indicates that the analysis is carried out at the level of a word or small group of words: first at the NER stage (in blue) and then during normalization (in green); finally, all mentions with normalized concept unique identifiers (CUIs) are aggregated at the “document level” (orange part). The sets of aggregated CUIs per document predicted by the native French and translated English approaches are then compared to the manually annotated gold standard.

Figure 1. Overall objective of the method: translating plain text to the CUI codes of the UMLS Metathesaurus, document by document. CHEM: Chemicals & Drugs; CUI: concept unique identifier; DISO: Disorders; PROC: Procedures; UMLS: United Medical Language System.



In many languages other than English, efforts remain to be made to obtain such results, notably due to a much smaller quantity of accessible data [4]. In this context, our work explores the relevance of a translation step for the recognition and normalization of medical concepts in French biomedical documents. We compared 2 methods: (1) a native French approach where only annotated documents and resources in French are used and (2) a translation-based approach where documents are translated into English, in order to take advantage of existing tools and resources for this language that would allow the extraction of concepts mentioned in unpublished French texts without new training data (zero-shot), as proposed in van Mulligen et al [5].

We evaluated and discussed the results on several French biomedical corpora, including a new set of 42 annotated hospitalization reports with 4 entity groups. We evaluated the normalization task at the document level, in order to avoid a cross-language alignment step at evaluation time, which would add a potential level of error and thus make the results more difficult to interpret (see word alignment in Gao and Vogel [6] and Vogel et al [7]). This normalization was carried out by mapping all terms to their CUI in the UMLS Metathesaurus [1]. Figure 1 summarizes these various stages, from the raw French text and the translated English text to the aggregation and comparison of CUIs at the document level. Our code is available on GitHub [8].

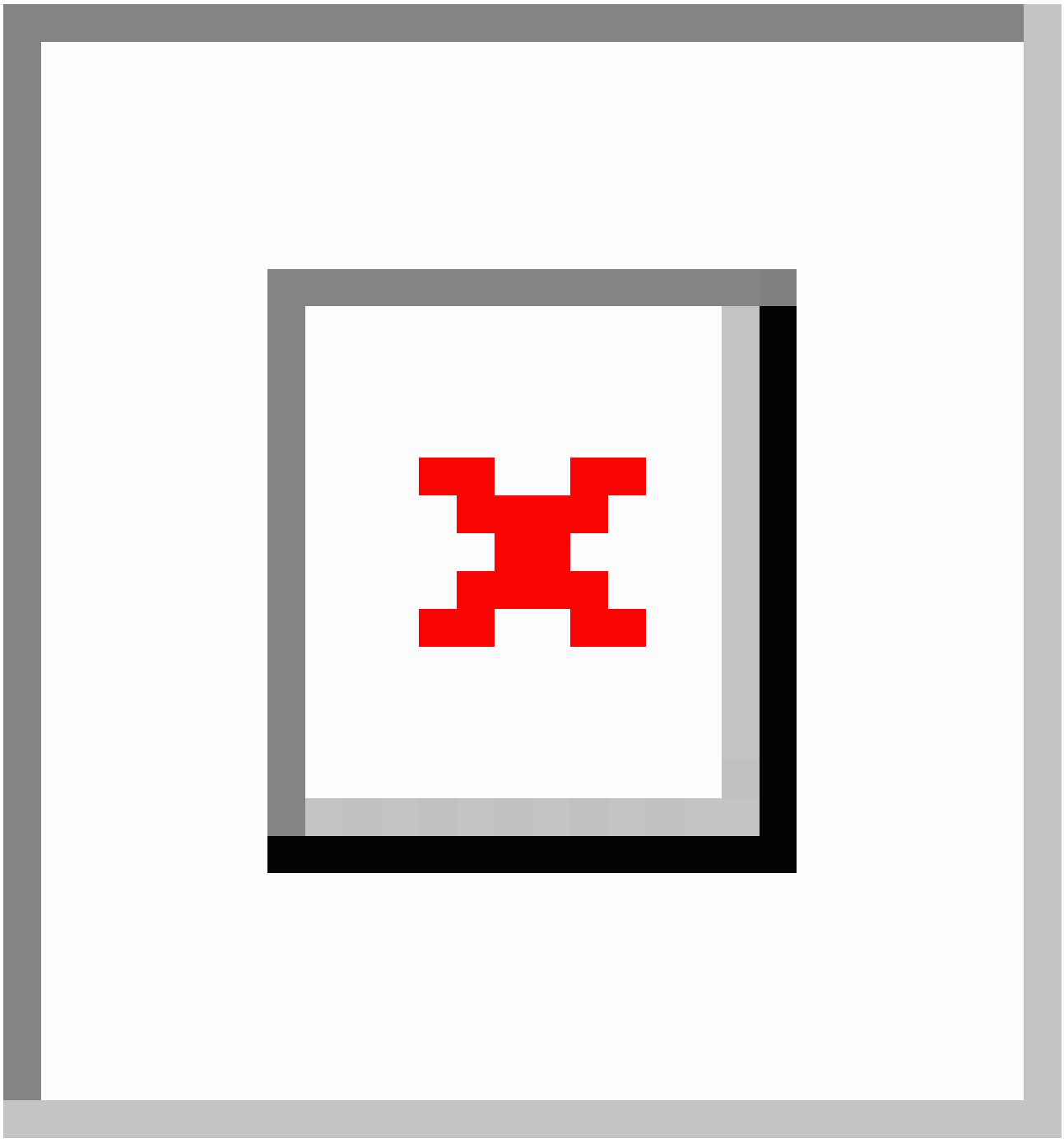
The various stages of our algorithms rely heavily on transformers language models [2]. These models currently represent the state of the art for many NLP tasks, such as machine translation, NER, classification, and text normalization (also known as entity binding). Once trained, these models can represent any specific language, such as biomedical or legal. The power of these models comes from their neural architecture but also largely depends on the amount of data they are trained on. In the biomedical field, 2 main types of data are available: public articles (eg PubMed) and clinical electronic medical record databases (eg MIMIC-III [9]), and the most powerful models are, for example, BioBERT [10], which has been trained on the whole of PubMed in English, and ClinicalBERT [11], which has been trained on PubMed and MIMIC-III. In French, the variety of models is less extensive, with CamemBERT [12] and FlauBERT [13] for the general domain and no specific model available for the biomedical domain.

In Figure 2, axis 1 (green axis on the left) corresponds to the native French branch with a NER step based on a FastText model trained from scratch on French clinical notes and a CamemBERT model. A multilingual Bidirectional Encoder Representations From Transformers (BERT) model was then used for the normalization step, with 2 models tested: a deep multilingual normalization model [14] and CODER [15] with the full version. Axes 2.1 and 2.2 (the 2 purple axes on the right) correspond to the translated English branches, with a first translation step performed by the OPUS-MT-FR-EN model [16] for both. Axis 2.1 (left) was conducted with decoupled NER and normalization steps: FastText trained from PubMed and MIMIC-III [17] for NER, and deep multilingual normalization [14] or CODER [15] with the English version for normalization. Axis 2.2 (right) used a single system for the NER and normalization stages: MedCAT [18].

In addition to particularly powerful English-language pretrained models, universal biomedical terminologies (ie, the UMLS Metathesaurus) also contain many more English terms than other languages. For example, the UMLS Metathesaurus [1] contains at least 10 times more English terms than French terms, which may enable rule-based models to perform better in English. As mentioned above, each reference concept in the UMLS Metathesaurus [1] is assigned a CUI, associated with a set of synonyms, possibly in several languages, and a semantic group, such as *Disorders*, *Chemicals & Drugs*, *Procedure*, *Anatomy*, etc.

In parallel, the performance of machine translation has also improved thanks to the same type of transformer-based language models, and recent years have seen the emergence of high-quality machine translations, such as OPUS-MT developed by Tiedemann et al [16], Google Translate, and others. These 2 observations have led several research teams to add a translation step in order to analyze medical texts, for example, to extract relevant mentions in ultrasound reports [19,20] or in the case of the standardization of medical concepts [14,15,21]. Work in the general (nonmedical) domain has also focused on alignment between named entities in parallel bilingual texts [22,23].

Figure 2. Diagram of different experiments comparing French and English language models without and with intermediate translation steps. CHEM: Chemicals & Drugs; CUI: concept unique identifier; DEVI: Devices; DISO: Disorders; EHR: electronic health record; EN: English; FR: French; FT: fine-tuned; PROC: Procedures; UMLS: United Medical Language System.



Methods

Approaches

Overview

Figure 2 shows the main approaches and models used in our study. We explored 1 "native French approach axis" (axis 1 in Figure 2), based on French linguistic models learned from and applied to French annotated data, and 2 "translated English approach axes" (axes 2.1 and 2.2), based on a translation step and concept extraction tools in English. We compared the performance of all axes with the average of the document-level CUI prediction precisions for all documents.

Native French Approach

Axis 1 consisted of 2 stages: a NER stage and a normalization stage. For the NER stage, we used the nested NER algorithm. Next, a normalization step was performed by 2 different algorithms: a deep multilingual normalization model [14] and CODER [15] with the *CODER all* version.

Translated-English Approach

First, axes 2.1 and 2.2 consisted of a translation step, performed by the state-of-the-art OPUS-MT-FR-EN [16] or Google

Translate algorithm. Second, similar to axis 1, axis 2.1 was based on a NER step and a normalization step. The NER step was performed by the same algorithm but trained on the National NLP Clinical Challenges (N2C2) 2019 data set [24] without manual annotation realignment; for the normalization step, we used the same deep multilingual algorithm [14] and the English version of CODER [15] based on a BioBERT model [10]. This axis allows us to compare 2 methods whose difference lies solely in the translation step.

Axis 2.2 was based on the MedCAT [18] algorithm, which performs NER and normalization simultaneously. In this case, we compared the native French method with a state-of-the-art, ready-to-use English system, which is not available in French.

Data Sets

For all our experiments, we chose to focus on 4 semantic groups of the UMLS Metathesaurus [1]: *Chemical & Drugs* ("CHEM"); *Devices* ("DEVI"), corresponding to medical devices such as pacemakers, catheters, etc; *Disorders* ("DISO"), corresponding to all signs, symptoms, results (eg, positive or negative results of biological tests), and diseases; and *Procedures* ("PROC"), corresponding to all diagnostic and therapeutic procedures such as imaging, biological tests, operative procedures, etc, as well as the corresponding number of documents.

Table 1 shows the data sets used for all our experiments and the corresponding number of documents. First, 2 French data sets were used for the final evaluation, as well as for training the axis-1 models. QUAERO is a freely available corpus [25] based on pharmacological notes with 2 subcorpora: MEDLINE (short sentences from PubMed abstracts) and EMEA (drug package inserts). We also annotated a new data set of real-life clinical notes from the Assistance Publique Hôpitaux de Paris data warehouse, described in Section S1 in Multimedia Appendix 1.

Table . Overview of all data sets used. When a data set is used for both training and testing, 80% of the data set is used for training and 20% is used for testing. Thus, for the EMEA data set, 30 documents were used for training and 8 for testing, 34 French notes were used for training and 8 for testing, and so on.

Variables	Languages and data sets						
	French			English		English and French	
	QUAERO [25]		French notes	N2C2[a] 2019 [24]	Mantra [26]	WMT[b] 2016 [27]	WMT 2019 [28]
	EMEA	MEDLINE					
Type	Drug notices	MEDLINE titles	French notes	English notes	Drug notices and MEDLINE titles	PubMed abstracts	PubMed abstracts
Size (documents), n	38	2514	42	100	200	>600,000 sent	6542
Use							
Train NER[c]	✓	✓	✓	✓			
Test NER	✓	✓	✓	✓			
Normalization	✓	✓	✓	✓			
Test MedCAT				✓	✓		
Translation (fine-tuning)						✓	✓
Translation (test)						✓	

[a]N2C2: National Natural Language Processing Clinical Challenges.

[b]WMT: Workshop on Machine Translation.

[c]NER: named entity recognition.

Second, we used the N2C2 2019 corpus [24] with annotated CUIs, on which we automatically added semantic group information from the UMLS Metathesaurus [1], to train the axis-2.1 system and evaluate the NER and English normalization algorithms. We also used the Mantra data set [26], a multilingual reference corpus for biomedical concept recognition.

Finally, we refined and tested the translation algorithms on the Workshop on Machine Translation biomedical corpora of 2016 [27] and 2019 [28]. A detailed description of the number of respective entities in the data sets can be found in Table S1 in Multimedia Appendix 1.

The annotation methods for the French corpus are detailed in Section S1 and Figure S1 in Multimedia Appendix 1. The distribution of entities for this annotation is detailed in Table S1 in Multimedia Appendix 1.

Translation

We used and compared 2 main algorithms for the translation step: the OPUS-MT-FR-EN model [16], which we tested without and with *fine-tuning* on the 2 biomedical translation corpora of 2016 and 2019 [27,28], and Google Translate as a comparison model.

NER Algorithm

For this step, we used the algorithm of Wajsbürt [29] described in Gérardin et al [30]. This model is based on the representation of a BERT transformer [3] and calculates the scores of all possible concepts to be predicted in the text. The extracted concepts are delimited by 3 values: start, end, and label. More precisely, the encoding of the text corresponds to the last 4 layers of BERT, FastText integration, and a max-pool Char-CNN [31] representation of the word. The decoding step is then performed by a 3-layer long short-term memory [32] with learning weights [33], similar to the method in Yu et al [34]. A sigmoid function was added to the vertex. Values (start, end, and label) with a score greater than 0.5 were retained for prediction. The loss function was a binary cross-entropy, and we used the Adam optimizer [35].

In our experiments, for the native French axis (axis 1 in Figure 2), the pretrained embeddings used to train the model were based on a FastText model [36], trained from scratch on 5 gigabytes of clinical text, and a CamemBERT-large model [12] *fine-tuned* on this same data set. For English axis 2.1, the pretrained models were BioWordVec [17] and ClinicalBERT [11].

Normalization Algorithms

Overview

This stage of our experiments was essential for comparing a method in native French and one translated into English, and it consisted of matching each mention extracted from the text to its associated CUI in the UMLS Metathesaurus [1]. We compared 3 models for this step, described below: the deep multilingual normalization algorithm developed by Wajsbürt et al [14]; CODER [15]; and the MedCAT [18] model, which performs both NER and normalization.

These 3 models require no training data set other than the UMLS Metathesaurus.

Deep Multilingual Normalization

This algorithm by Wajsbürt et al [14] considers the normalization task as a highly multiclass classification problem with cosine similarity and a softmax function as the last layer. The model is based on contextual integration, using the pretrained multilingual BERT model [3], and works in 2 steps. In the first step, the BERT model is fine-tuned and the French UMLS terms and their corresponding English synonyms are learned. Then, in the second step, the BERT model is frozen and the representation of all English-only terms (ie, those present only in English in the UMLS Metathesaurus [1]) is learned. The same training is used for the native French and translated English approaches. This model was trained with the 2021 version of the UMLS Metathesaurus [1], corresponding to the version used for annotating the French corpus. The model was thus trained on over 4 million concepts corresponding to 2 million CUIs.

CODER

The CODER algorithm [15] was developed by contrastive learning on the basis of the medical knowledge graph of the UMLS Metathesaurus [1], with concept similarities being calculated from the representation of terms and relations in this knowledge graph. Contrastive learning is used to learn embeddings through multisimilarity loss [37]. The authors have developed 2 versions: a multilingual version based on the multilingual BERT [3] and an English version based on the pretrained BioBERT model [10]. We used the multilingual version for axis 1 (native French approach) and the English version for axis 2.1. Both types of this model (*CODER all* and *CODER en*) were trained with the 2020 version of UMLS (publicly available models). *CODER all* [15] was trained on over 4 million concepts corresponding to 2 million CUIs, and *CODER en* was trained on over 3 million terms and 2 million CUIs.

For the deep multilingual model and the CODER model, in order to improve performance in terms of accuracy, we chose to add semantic group information (ie, *Chemical & Drugs*, *Devices*, *Disorders*, and *Procedures*) to the model output: that is, from the first *k* CUIs chosen from a mention, we selected the first from the corresponding group.

The MedCAT algorithm is described in detail in Section S1 in Multimedia Appendix 1.

Ethical Considerations

The study and its experimental protocol were approved by the Assistance Publique Hôpitaux de Paris Scientific and Ethical Committee (IRB00011591, decision CSE 20-0093). Patients were informed that their electronic health record information could be reused after an anonymization process, and those who objected to the reuse of their data were excluded. All methods were applied in accordance with the relevant guidelines (*Commission nationale de l'informatique et des libertés* reference methodology MR-004 [38]).

Results

The sections below present the performance results for each stage. The N2C2 2019 challenge corpus [24] enabled us to evaluate the performance of our English models on clinical data, and the Biomedical Translation 2016 shared task [27] allowed us to evaluate our translation performance on biomedical data with a BLEU score [39].

NER Performances

To be able to compare our approaches in native French and translated English, we used the same NER model, trained and tested on each of the data sets described above. Table 2 shows the corresponding results. Overall F_1-scores were similar across data sets: from 0.72 to 0.77.

Table . Named entity recognition (NER) performance for each model. For all experiments, we used the same NER algorithm but with different pretrained models. The best performance values are italicized.

Groups	Data sets and models								
	EMEA test, with FastText*[a] and Camem-BERT-FT [12]			French notes, with FastText* and Camem-BERT-FT			N2C2[b] 2019 test, with BioWordVec [17] and ClinicalBERT [11]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
CHEM[c]	0.80	0.83	0.82	0.84	0.88	0.86	0.87	0.85	0.86
DEVI[d]	0.42	0.81	0.55	0.00	0.00	0.00	0.58	0.51	0.54
DISO[e]	0.54	0.63	0.59	0.67	0.65	0.66	0.74	0.72	0.73
PROC[f]	0.73	0.78	0.74	0.78	0.72	0.75	0.80	0.78	0.79
Overall	*0.71*	*0.77*	*0.74*	*0.73*	*0.71*	*0.72*	*0.78*	*0.76*	*0.77*

[a]FastText* corresponds to a FastText model [36] trained from scratch on our clinical data set.

[b]N2C2: National Natural Language Processing Clinical Challenges.

[c]CHEM: Chemical & Drugs.

[d]DEVI: Devices.

[e]DISO: Disorders.

[f]PROC: Procedures.

Normalization Performances

This section presents only the normalization performance based on the gold standard's entity mentions, without the intermediate steps. The results are summarized in Table 3. The deep multilingual algorithm performed better for all corpora tested, with an improvement in F_1-score from +0.6 to +0.11. By way of comparison, the winning team of the 2019 N2C2 had achieved an accuracy of 0.85 using the N2C2 data set directly to train their algorithm [24]. In our context of comparing algorithms between 2 languages, the normalization algorithms were not trained on data other than the UMLS Metathesaurus. MedCAT's performance (shown in Table S2 in Multimedia Appendix 1) cannot be directly compared with that of other models, as this method performed both NER and normalization in a single step. However, we note that this algorithm performed as well as axis 2.1 in terms of overall performance, as shown in Table 4.

Table . Performance of the normalization step. Model results were calculated from the annotated data sets, focusing on the 4 semantic groups of interest: *Chemical & Drugs*, *Devices*, *Disorders*, and *Procedures*. The best performance values are italicized.

Algorithms	Data set models		
	EMEA test	French notes	N2C2[a] 2019 test
Deep multilingual normalization	*0.65*	*0.57*	*0.74*
CODER all	0.58	0.51	—[b]
CODER en	—	—	0.63

[a]N2C2: National Natural Language Processing Clinical Challenges.

[b]Not applicable.

Table . Overall performances. The normalization step was performed by the deep multilingual model and the translation was performed by the OPUS-MT-FR-EN FT model. The best performance values are italicized.

Methods	EMEA test			French notes		
	Precision	Recall	F_1-score (95% CI)	Precision	Recall	F_1-score (95% CI)
Axis 1 (French NER[a]+normalization)	0.63	0.60	*0.61 (0.53-0.65)*	0.49	0.53	*0.51 (0.47-0.55)*
Axis 2.1 (Translation+NER+normalization)	0.53	0.40	0.45 (0.38-0.51)	0.41	0.38	0.39 (0.34-0.44)
Axis 2.2 (Translation+MedCAT [18])	0.53	0.46	0.49 (0.38-0.54)	0.38	0.38	0.38 (0.36-0.40)

[a]NER: named entity recognition.

Translation Performances

For both translation models, the respective BLEU scores [39] were calculated on the shared 2016 Biomedical Translation Task [27]. The chosen BLEU algorithm was the weighted geometric mean of the n-gram precisions per sentence.

A fine-tuned version of OPUS-MT-FR-EN [16] was also tested on the 2016 and 2019 Biomedical Translation shared tasks. For fine-tuning, we used the following hyperparameters: a maximum sequence length of 128 (mainly for computational memory reasons), a learning rate of 2×10^{-5}, and a weight decay of 0.01, and we varied the number of epochs up to 15 epochs (the error function curve stops decaying after 10 epochs). The Google Translate model could not be used for our clinical score experiments for reasons of confidentiality.

Table 5 presents the BLEU scores for the 3 models, showing that fine-tuning the OPUS-MT-FR-EN model [16] on biomedical data sets gave the best results, with a BLEU score [39] of 0.51. This was the model used to calculate the overall performance of axes 2.1 and 2.2.

Table . Translation performances: BLEU scores of the translation models. The best performance value is italicized.

Models	WMT[a] Biomed 2016 test
Google Translate	0.42
OPUS-MT-FR-EN	0.31
OPUS-MT-FR-EN FT[b]	*0.51*

[a]WMT: Workshop on Machine Translation.

[b]OPUS-MT-FR-EN FT corresponds to the OPUS-MT-FR-EN model [16] *fine-tuned* on biomedical translated corpus from the WMT Biomedical Translation Tasks in 2016 [27] and 2019 [28].

Overall Performances From Raw Text to CUI Predictions

This section presents the overall performance of the 3 axes, in an end-to-end pipeline. For axis 2, the results are those obtained with the best normalization algorithm (presented in Table 3). The model used for translation is the OPUS-MT-FR-EN [16] fine-tuned model. The results are presented in Table 4, with the best results obtained by the native French approach on the EMEA corpus [25] and French clinical notes. The 95% CIs were calculated using the empirical bootstrap method [40].

Discussion

Principal Findings

In this paper, we compared 2 approaches for extracting medical concepts from clinical notes: a French approach based on a French language model and a translated English approach, where we compared 2 state-of-the-art English biomedical language models, after a translation step. The main advantages of our experiment are that it is reproducible and that we were able to analyze the performance of each step of the algorithm: NER, normalization, and translation, and to test several models for each step.

The Quality of the Translation Is Not Sufficient

We showed that the native French approach outperformed the 2 translated English approaches, even with a small French training data set. This analysis confirms that, where possible, an annotated data set improves feature extraction. The evaluation of each intermediate step showed that the performance of each module was similar in French and English. We can therefore conclude that it is rather the translation phase itself that is of insufficient quality to allow the use of English as a proxy without a loss of performance. This is confirmed by the translation performance calculations, where the calculated BLEU scores were relatively low, although improved by a fine-tuning step.

In conclusion, although translation is commonly used for entity extraction or term normalization in languages other than English

[5,20,41-43], due to the availability of turnkey models that do not require additional annotation by a clinician, we showed that this induces a significant performance loss.

Commercial application programming interface–based translation services could not be used for our task due to data confidentiality issues. However, the OPUS-MT model is considered state of the art, it is adjustable to domain-specific data, and the translation results presented in Table 5 confirm the absence of performance difference between this model and the Google Translate model.

Although our experiments were carried out on a single language, the French-English pair is one of the best performers in recent translation benchmarks [16]. Other languages are unlikely to produce significantly better results.

Error Analysis

In these experiments, the overall results may appear low, but the task is still complex, especially because the UMLS Metathesaurus [1] contains many synonyms with different CUIs. To better understand this, we performed an error analysis on the normalization task only, as shown in Table S3 in Multimedia Appendix 1, with a physician's evaluation, on a sample of 100 errors for both models. We calculated that 24% (24/100) and 39% (39/100) of the terms found by the deep normalization algorithm [14] and CODER [15], respectively, were in fact synonyms but had 2 different UMLS CUIs. This highlights the difficulty of achieving normalization on the UMLS Metathesaurus. The UMLS Metathesaurus indeed groups together numerous terminologies whose mapping between terms is often imperfect, implying that certain synonyms, as shown here, do not have the same CUI, as pointed out by Cimino [44] and Jiménez-Ruiz et al [45]. For example, "cardiac ultrasound" has the CUI of C1655737, whereas "echocardiography" has another CUI of C0013516; similarly, "H/O: thromboembolism" has a CUI of C0455533, whereas "history of thromboembolism" has a CUI of C1997787, and so on.

Moreover, to be more precise, each axis had its own errors: overall, the errors in axis 2 were essentially due to the loss of information in translation. One notable error was literal translation: for example, "dispersed lupus erythematous" instead of "systemic lupus erythematosus," or "crepitant" instead of "crackles." This loss of translation led to more errors in the extraction of named entities.

In addition to the loss of translation information, axis 2.1 was also penalized by the NER step, due to the difference between the training set (N2C2 notes) and the test set (the translated French notes; the aim being to compare the performance of English-language turnkey models with the performance of French-language models from an annotated set). Axis 2.1, for example, omitted the names of certain drugs more often.

Finally, both axes were penalized by abbreviations. These were often badly translated (for example, the abbreviation "MFIU" for "mort foetale in utero," meaning "intrauterine fetal death," was not translated), which penalized axis 2. Nevertheless, if they were indeed extracted by NER steps in axis 1, they were not correctly normalized due to the absence of a corresponding CUI in the UMLS Metathesaurus.

Limitations

This work has several limitations. First, the actual French clinical notes contained very few terms in the *Devices* semantic group, which prevented the NER algorithm from finding them in the test data set. However, this drawback, which penalized the native French approach, still allowed us to draw a conclusion for the results. Furthermore, in this study, we did not take into account attributes of the extracted terms such as negation, hypothetical attribute, or belonging to a person other than the patient for comparison purposes, as the QUAERO [25] and N2C2 2019 [24] data sets did not have this labeled information.

Acknowledgments

The authors would like to thank the Assistance Publique Hôpitaux de Paris (AP-HP) data warehouse, which provided the data and the computing power to carry out this study under good conditions. We wish to thank all the medical colleges, including internal medicine, rheumatology, dermatology, nephrology, pneumology, hepato-gastroenterology, hematology, endocrinology, gynecology, infectiology, cardiology, oncology, emergency, and intensive care units, that gave their permission for the use of the clinical data.

Data Availability

The data sets analyzed as part of this study are not accessible to the public due to the confidentiality of data from patient files, even after deidentification. However, access to raw data from the Assistance Publique Hôpitaux de Paris (AP-HP) data warehouse can be granted by following the procedure described on its website [46]: by contacting the ethical and scientific committee at secretariat.cse@aphp.fr. Prior validation of access by the local institutional review committee is required. In the case of non-APHP researchers, a collaboration contract must also be signed.

Authors' Contributions

CG contributed to conceptualization, data curation, formal analysis, investigation, methodology, software, validation, original drafting, writing—original version, and writing—revision and editing the manuscript. YX contributed to investigation, methodology, software, and validation. PW contributed to investigation, software, and revision of the manuscript. FC contributed to conceptualization, methodology, project administration, supervision, writing—original version, and writing—revision and editing

of the manuscript. XT contributed to conceptualization, formal analysis, methodology, writing—original version, and writing—revision and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed description of the data sets, an example of the clinical notes annotation, French corpus annotation, MedCAT performances, and error analysis.
[DOCX File, 154 KB - medinform_v12i1e49607_app1.docx]

References

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(suppl 1):D267-D270. [doi: 10.1093/nar/gkh061] [Medline: 14681409]
2. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, et al, editors. Advances in Neural Information Processing Systems 30 (NIPS 2017) 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [accessed 2024-03-15]
3. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. 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): Association for Computational Linguistics; 2019:4171-4186. [doi: 10.18653/v1/N19-1423]
4. 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]
5. van Mulligen EM, Afzal Z, Akhondi SA, Vo D, Kors JA. Erasmus MC at CLEF Ehealth 2016: concept recognition and coding in French texts. In: Balog K, Cappellato L, Ferro N, Macdonald C, editors. Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum CEUR Workshop Proceedings, Vol 1609: CEUR-WS.org; 2016:171-178 URL: https://ceur-ws.org/Vol-1609/16090171.pdf [accessed 2024-03-15]
6. Gao Q, Vogel S. Parallel Implementations of word alignment tool. In: Cohen KB, Carpenter B, editors. SETQA-NLP '08: Software Engineering, Testing, and Quality Assurance for Natural Language Processing: Association for Computational Linguistics; 2008:49-57. [doi: 10.5555/1622110.1622119]
7. Vogel S, Ney H, Tillmann C. HMM-based word alignment in statistical translation. In: COLING '96: Proceedings of the 16th Conference on Computational Linguistics - Volume 2: Association for Computational Linguistics; 1996:836-841. [doi: 10.3115/993268.993313]
8. ChristelDG/biomed_translation. GitHub. URL: https://github.com/ChristelDG/biomed_translation [accessed 2024-03-15]
9. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
10. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
11. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on Apr 10, 2019. [doi: 10.48550/arXiv.1904.05342]
12. Martin L, Muller B, Ortiz Suárez PJ, et al. CamemBERT: a tasty French language model. In: Kurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2020:7203-7219. [doi: 10.18653/v1/2020.acl-main.645]
13. Le H, Vial L, Frej J, et al. FlauBERT: unsupervised language model pre-training for French. In: Calzolari N, Béchet F, Blanche P, et al, editors. Proceedings of the Twelfth Language Resources and Evaluation Conference: European Language Resources Association; 2020:2479-2490 URL: https://aclanthology.org/2020.lrec-1.302 [accessed 2024-03-15]
14. Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. J Biomed Inform 2021 Feb;114:103684. [doi: 10.1016/j.jbi.2021.103684] [Medline: 33450387]
15. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform 2022 Feb;126:103983. [doi: 10.1016/j.jbi.2021.103983] [Medline: 34990838]
16. Tiedemann J, Thottingal S. OPUS-MT - building open translation services for the world. In: Martins A, Moniz H, Fumega S, et al, editors. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation: European Association for Machine Translation; 2020:479-480 URL: https://aclanthology.org/2020.eamt-1.61 [accessed 2024-03-15]
17. 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]
18. Kraljevic Z, Bean D, Mascio A, et al. MedCAT -- medical concept annotation tool. arXiv. Preprint posted online on Dec 18, 2019. [doi: 10.48550/arXiv.1912.10166]

19. Campos L, Pedro V, Couto F. Impact of translation on named-entity recognition in radiology texts. Database (Oxford) 2017 Jan 1;2017(2017):bax064. [doi: 10.1093/database/bax064] [Medline: 29220455]
20. Suarez-Paniagua V, Dong H, Casey A. A multi-BERT hybrid system for named entity recognition in Spanish radiology reports. In: Faggioli G, Ferro N, Joly A, Maistro M, Piroi F, editors. Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, Vol 2936: CEUR-WS.org; 2021:846-856 URL: https://ceur-ws.org/Vol-2936/paper-70.pdf [accessed 2024-03-15]
21. Perez-Miguel N, Cuadros M, Rigau G. Biomedical term normalization of EHRs with UMLS. In: Calzolari N, Choukri K, Cieri C, et al, editors. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018): European Language Resources Association (ELRA); 2018:2045-2051 URL: https://aclanthology.org/L18-1322 [accessed 2024-03-15]
22. Chen Y, Zong C, Su KYS. On jointly recognizing and aligning bilingual named entities. In: Hajič J, Carberry S, Clark S, Nivre J, editors. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2010:631-639 URL: https://aclanthology.org/P10-1065 [accessed 2024-03-15]
23. Chen Y, Zong C, Su KYS. A joint model to identify and align bilingual named entities. Comput Linguist 2013 Jun 1;39(2):229-266. [doi: 10.1162/COLI_a_00122]
24. Henry S, Wang Y, Shen F, Uzuner O. The 2019 National Natural Language Processing (NLP) Clinical Challenges (N2C2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records. J Am Med Inform Assoc 2020 Oct 1;27(10):1529-1537. [doi: 10.1093/jamia/ocaa106] [Medline: 32968800]
25. Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French medical corpus: a resource for medical entity recognition and normalization. Presented at: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing - BioTextM2014; May 26-31, 2014; Reykjavik, Iceland p. 24-30 URL: https://perso.limsi.fr/pz/FTPapiers/Neveol_BIOTEXTM2014.pdf [accessed 2024-03-15]
26. Kors JA, Clematide S, Akhondi SA, van Mulligen EM, Rebholz-Schuhmann D. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. J Am Med Inform Assoc 2015 Sep;22(5):948-956. [doi: 10.1093/jamia/ocv037] [Medline: 25948699]
27. Bojar O, Chatterjee R, Federmann C. Findings of the 2016 Conference on Machine Translation. In: Bojar O, Buck C, Chatterjee R, et al, editors. Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers: Association for Computational Linguistics; 2016:131-198. [doi: 10.18653/v1/W16-2301]
28. Bawden R, Bretonnel Cohen K, Grozea C, et al. Findings of the WMT 2019 Biomedical Translation Shared Task: evaluation for MEDLINE abstracts and biomedical terminologies. In: Bojar O, Chatterjee R, Federmann C, et al, editors. Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2): Association for Computational Linguistics; 2019:29-53. [doi: 10.18653/v1/W19-5403]
29. Wajsbürt P. Extraction and Normalization of Simple and Structured Entities in Medical Documents [thesis].: Sorbonne Université; 2021 Dec URL: https://theses.hal.science/THESES-SU/tel-03624928v1 [accessed 2024-03-15]
30. Gérardin C, Wajsbürt P, Vaillant P, Bellamine A, Carrat F, Tannier X. Multilabel classification of medical concepts for patient clinical profile identification. Artif Intell Med 2022 Jun;128:102311. [doi: 10.1016/j.artmed.2022.102311] [Medline: 35534148]
31. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Knight K, Nenkova A, Rambow O, editors. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Association for Computational Linguistics; 2016:260-270. [doi: 10.18653/v1/N16-1030]
32. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
33. Kim J, El-Khamy M, Lee J. Residual LSTM: design of a deep recurrent architecture for distant speech recognition. 2017 Presented at: Interspeech 2017; Aug 20-24, 2017; Stockholm, Sweden p. 1591-1595. [doi: 10.21437/Interspeech.2017-477]
34. Yu J, Bohnet B, Poesio M. Named entity recognition as dependency parsing. In: Jurafsky D, Chai J, Schulter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Association for Computational Linguistics; 2020:6470-6476. [doi: 10.18653/v1/2020.acl-main.577]
35. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. Preprint posted online on Dec 22, 2014. [doi: 10.48550/arXiv.1412.6980]
36. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017 Dec 1;5:135-146. [doi: 10.1162/tacl_a_00051]
37. Wang X, Han X, Huang W, Dong D, Scott MR. Multi-similarity loss with general pair weighting for deep metric learning. 2019 Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 15-20, 2019; Long Beach, CA p. 5017-5025. [doi: 10.1109/CVPR.2019.00516]
38. CNIL (Commission Nationale de l'Informatique et des Libertés). URL: https://www.cnil.fr/en/home [accessed 2024-03-15]
39. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics: Association for Computational Linguistics; 2002:311-318. [doi: 10.3115/1073083.1073135]

40. Dekking FM, Kraaikamp C, Lopuhaa HP, Meester LE. A Modern Introduction to Probability and Statistics: Understanding Why and How: Springer Nature; 2007.
41. Cotik V, Rodríguez H, Vivaldi J. Spanish named entity recognition in the biomedical domain. In: Lossio-Ventura J, Muñante D, Alatrista-Salas H, editors. Information Management and Big Data. SIMBig 2018. Communications in Computer and Information Science, vol 898: Springer:233-248. [doi: 10.1007/978-3-030-11680-4]
42. Hellrich J, Hahn U. Enhancing multilingual biomedical terminologies via machine translation from parallel corpora. In: Métais E, Roche M, Teisseire M, editors. Natural Language Processing and Information Systems. NLDB 2014. Lecture Notes in Computer Science, vol 8455: Springer; 2014:9-20. [doi: 10.1007/978-3-319-07983-7_2]
43. Attardi G, Buzzelli A, Sartiano D. Machine translation for entity recognition across languages in BIOMEDICAL documents. In: Forner P, Navigli R, Tufis D, Ferro N, editors. Working Notes for CLEF 2013 Conference. CEUR Workshop Proceedings, Vol 1179: CEUR-WS.org; 2013. URL: https://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFER-AttardiEt2013.pdf [accessed 2024-03-15]
44. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51. [doi: 10.1136/jamia.1998.0050041] [Medline: 9452984]
45. Jiménez-Ruiz E, Grau BC, Horrocks I, Berlanga R. Logic-based assessment of the compatibility of UMLS ontology sources. J Biomed Semantics 2011 Mar 7;2 Suppl 1(Suppl 1):S2. [doi: 10.1186/2041-1480-2-S1-S2] [Medline: 21388571]
46. Assistance Publique Hôpitaux de Paris. URL: www.eds.aphp.fr [accessed 2024-03-18]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
CUI: concept unique identifier
N2C2: National Natural Language Processing Clinical Challenges
NER: named entity recognition
NLP: natural language processing
UMLS: United Medical Language System

Edited by C Lovis; submitted 03.06.23; peer-reviewed by L Modersohn, M Torii; revised version received 07.01.24; accepted 10.01.24; published 04.04.24.

Please cite as:
Gérardin C, Xiong Y, Wajsbürt P, Carrat F, Tannier X
Impact of Translation on Biomedical Information Extraction: Experiment on Real-Life Clinical Notes
JMIR Med Inform 2024;12:e49607
URL: https://medinform.jmir.org/2024/1/e49607
doi:10.2196/49607

Original Paper

Additional Value From Free-Text Diagnoses in Electronic Health Records: Hybrid Dictionary and Machine Learning Classification Study

Tarun Mehra[1], MD; Tobias Wekhof[2], PhD; Dagmar Iris Keller[3,4], MD

[1]Department for Medical Oncology and Hematology, University Hospital of Zurich, Zurich, Switzerland

[2]Center of Economic Research, ETH Zurich, Zurich, Switzerland

[3]Faculty of Medicine, University of Zurich, Zurich, Switzerland

[4]Emergency Department, University Hospital of Zurich, Zurich, Switzerland

Corresponding Author:
Tarun Mehra, MD
Department for Medical Oncology and Hematology
University Hospital of Zurich
Rämistrasse 100
Zurich, 8091
Switzerland
Phone: 41 44255 ext 1111
Email: tarun.mehra@usz.ch

Abstract

Background: Physicians are hesitant to forgo the opportunity of entering unstructured clinical notes for structured data entry in electronic health records. Does free text increase informational value in comparison with structured data?

Objective: This study aims to compare information from unstructured text-based chief complaints harvested and processed by a natural language processing (NLP) algorithm with clinician-entered structured diagnoses in terms of their potential utility for automated improvement of patient workflows.

Methods: Electronic health records of 293,298 patient visits at the emergency department of a Swiss university hospital from January 2014 to October 2021 were analyzed. Using emergency department overcrowding as a case in point, we compared supervised NLP-based keyword dictionaries of symptom clusters from unstructured clinical notes and clinician-entered chief complaints from a structured drop-down menu with the following 2 outcomes: hospitalization and high Emergency Severity Index (ESI) score.

Results: Of 12 symptom clusters, the NLP cluster was substantial in predicting hospitalization in 11 (92%) clusters; 8 (67%) clusters remained significant even after controlling for the cluster of clinician-determined chief complaints in the model. All 12 NLP symptom clusters were significant in predicting a low ESI score, of which 9 (75%) remained significant when controlling for clinician-determined chief complaints. The correlation between NLP clusters and chief complaints was low (r=−0.04 to 0.6), indicating complementarity of information.

Conclusions: The NLP-derived features and clinicians' knowledge were complementary in explaining patient outcome heterogeneity. They can provide an efficient approach to patient flow management, for example, in an emergency medicine setting. We further demonstrated the feasibility of creating extensive and precise keyword dictionaries with NLP by medical experts without requiring programming knowledge. Using the dictionary, we could classify short and unstructured clinical texts into diagnostic categories defined by the clinician.

(JMIR Med Inform 2024;12:e49007) doi:10.2196/49007

KEYWORDS

electronic health records; free text; natural language processing; NLP; artificial intelligence; AI

Introduction

Organizational challenges, such as overcrowding in emergency departments (EDs), directly impact patient outcomes. The digitization of health records offers an opportunity to integrate artificial intelligence (AI) into patient management. However, health care workers often prefer to write unstructured text rather than entering structured data [1,2]. This raises the question of how future electronic health records (EHRs) should be designed: what additional value does free text provide?

We propose adding an additional dimension alongside the classic predictive task performed with text—inference to infer characteristics from text entries. Most studies using text analysis with patient records show promising results in predicting patient outcomes, such as in-hospital mortality, unplanned re-admission after 30 days, and prolonged length of hospital stay [3,4]. The benefits of unstructured text in EHRs for the improvement of prediction models have been demonstrated, as underscored by the extensive review by Seinen et al [5]. Indeed, 20% of the trials that were reported were conducted within a hospital ED environment. However, the analysis of the reported studies focused on demonstrating an improvement in predicting clinical outcomes, such as death or rehospitalization. We extend this approach by using the text not primarily to predict outcomes but to explain the correlation of patient subgroups with clinical outcomes. For instance, we show if certain symptoms documented in the ED triage are associated with a higher probability of an inpatient stay. Our results indicate that the information captured by clinical text-based notes is complementary to traditional structured data and can provide clinicians with valuable information about patients.

Overcrowding in the ED is an important case in point where AI supporting the optimization of patient workflows may substantially improve outcomes. It is a recognized challenge facing many EDs worldwide [6,7], adversely impacting patient outcomes [8]. These negative effects are evident during ED resource overload, such as during the COVID-19 pandemic [9]. More recently, senior public health officials in England have attributed up to 500 excess deaths per week during the recent winter months to delays caused by National Health Service capacity constraints [10,11]. Therefore, electronically enabled targeted patient selection could help speed up triage and reduce ED overcrowding. However, the optimal structure of EHRs remains controversial, particularly because clinicians tend to prefer the flexibility of entering unstructured text to structured data entry [12].

By comparing data extracted from 2 fields—1 derived from a structured drop-down menu indicating leading symptoms for ED admission and the other containing unstructured text—we can demonstrate that free text contains additional information beyond structured data and that these 2 types of data complement each other. With our semisupervised topic allocation method, we demonstrate the ability to capture more comprehensive information about a patient's symptom cluster compared with relying solely on a manually attributed single chief complaint. Moreover, we present a transparent approach for extracting topics from short clinical texts based on natural language processing (NLP)–supported annotated clinical libraries, which can be fed into predictive models. In addition to being transparent, our method is language independent and easy to implement for clinical researchers (although the dictionaries we constructed are in German, researchers can easily use our method to construct their own topic dictionaries in any language).

Our approach is based on constructing a dictionary with keywords that define a topic. In contrast to dictionary approaches, unsupervised topic models, such as the latent Dirichlet allocation [13], are often used. However, finding topics in short-text samples using these models is challenging [14]. Moreover, unsupervised models might not capture topics that are of interest to the researcher because these models differentiate between topics based on their statistical difference. For instance, it could be that latent Dirichlet allocation defines topics based on words about the age and gender of the patients because these are the most distinctive features. However, the researcher may be interested in the diagnosis, which is more challenging to classify.

In contrast, supervised machine learning methods require creating a manually classified training data set. The algorithm learns how to classify future data into topics based on the training set. When dealing with a high volume of topics, both human classification and the algorithm's training run the risk of creating noise. Similarly, regression approaches for supervised classifications are not suitable for many topics. Therefore, we chose a dictionary approach based on keywords. To facilitate the selection of the keywords, we developed a preselection of words based on a measure of their semantic similarity. As our presorting of words uses word embeddings, we consider our approach as a hybrid between dictionary- and machine learning–based approaches [15].

Our approach, combined with clinical notes, allow us to address 2 questions:

- What additional information does the free text provide on the patient being admitted compared with the suspected diagnosis from the drop-down menu?
- Could this additional information be useful for clinical or organizational purposes?

Methods

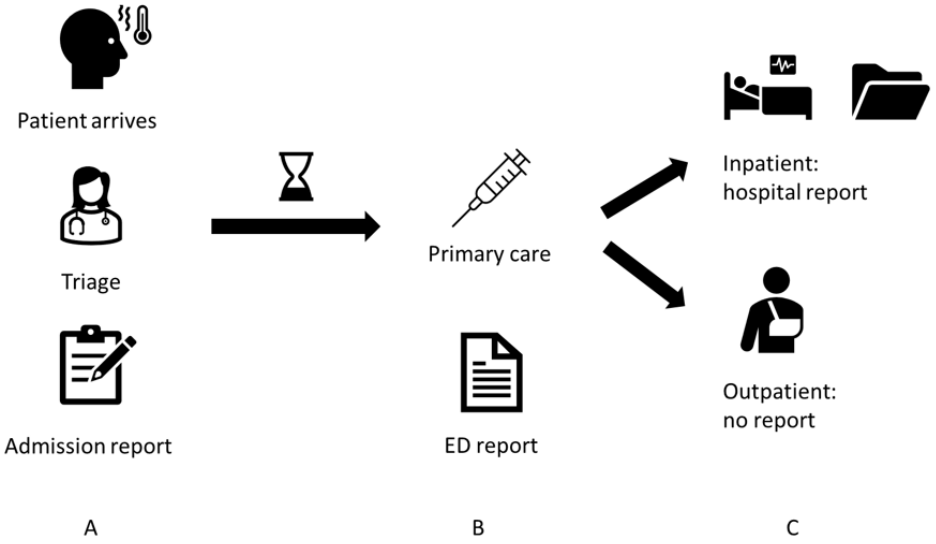
Data

We used data from the ED's admission report. Figure 1 provides a contextual representation of this data type in relation to patient flow and other documents associated with patients. In step 1, patients present themselves at the ED and are admitted in the system. A medical professional conducts the triage by quickly assessing the main symptoms and their severity using the Emergency Severity Index (ESI) score, resulting in an admission report. This report is for the internal patient management within the ED and contains basic patient information (age, gender, and so on) along with the chief complaints and symptoms.

After a waiting time (which depends on the triage score), the patient receives primary care from a medical professional, which

is documented in the ED report. The ED report summarizes the patient's entire stay at the ED and is issued at the end of the patient care from the ED. In the third step, the patient is either discharged into ambulatory care (which does not create any further documents) or is transferred to inpatient care, which results in the classic medical records.

Figure 1. Patient flow in emergency department (ED) and associated reports.



For our analysis, we used the first type of document: the internal ED admission report. Unlike the other types of documents, this report is issued before treatment and provides an opportunity to manage patient flow. Although the ED report from step 2 could also be used for inpatient management, this proves challenging in practice because inpatient care is very heterogeneous and depends on many factors, including different organizational structures in every hospital department. In contrast, the ED admission reports can be used for homogeneous organization within the ED.

Our initial data set contained 293,298 patient visits to the ED of the University Hospital of Zurich, Switzerland, from January 1, 2014, to October 31, 2021 (in German; received in the Excel [Microsoft Corporation] format). For each visit, the data set includes a short text from the triage with the patient's symptoms, along with our 2 outcomes of interest (triage score "ESI," which we further explain below, and type of discharge), basic patient characteristics (patient visit pseudo ID, age, gender, admission type [self, ambulance, or police], and admission reason [accident or illness]), ED organizational variables (average number of patients in ED; average patient waiting time; night, late, or early shift; and treating ED team [internal medicine, surgery, neurology, neurosurgery, or psychiatry]), and the visit's time stamp. The summary statistics of these variables are presented in Table 1.

After excluding cases with no records in the string variable "suspected diagnosis" on admission on which NLP analysis was to be performed, the data set comprised 256,329 (87.4%) of the initial data set of 293,298 patient visits. We only used 2019 to 2021 for comparison as these visits had a recorded chief complaint, reducing the data set to the final sample of 52,222 patient visits. Patients directly admitted to the shock room (ie, ESI score=1) were not considered in our analysis, as no additional triage was performed upon admission. The data structure of our analysis is summarized in Figure 2, and the recorded variables are presented in Textbox 1.

The ESI is an internationally established 5-level triage algorithm widely used in EDs and is based on the acuity as well as the resource intensity of anticipated emergency care, with level 1 denoting acute life-threatening conditions, such as massive trauma warranting immediate, life-saving care, and level 5 denoting non–time-critical conditions of low complexity [13]. Cases triaged as ESI 4 or 5 (approximately 16% of patients) are usually fast-tracked to specialized treatment rooms because the medical resources required to treat these patients are low, and thus, they can be managed in parallel by a dedicated team, which reduces ED congestion. ESI 2 or 3 typically require a more thorough workup. Hence, for the outcome variable "low ESI," we decided to set the cutoff at ESI<4, that is, patients with "low ESI" had been triaged with a score of 2 or 3. Furthermore, the data set included free-text fields (strings), namely, the suspected diagnosis at admission and the diagnosis at discharge.

In the admission process, the clinician performing triage records the patient's symptoms in written form in 2 to 3 sentences. The purpose of this free text is to preregister the patient in the ED and enable all team members to become aware of the impending clinical problems. To our knowledge, all the larger EDs in German-speaking countries with full EHR note the reason for admission in the form of a short, unstructured text upon notification of a pending ED admission.

From May 28, 2019, onward, the symptoms were additionally recorded as so-called chief complaints from a drop-down menu (ordinal variable). The difference between the free text and the chief complaint was that the chief complaint was a fixed category selected from a drop-down menu and was primarily intended to serve administrative and statistical purposes, that is, to allow for post hoc analysis of the patient composition of the ED.

During the entire study period, the list of chief complaints (n=99) varied over time or contained doublets, which we grouped into 58 symptom topics. For patient visits with a

selected chief complaint from the drop-down option "Diverse," it was unclear if a leading symptom had been attributed at triage; hence, we did not include them in the list of chief complaints (referred to as lead symptoms [LS]). Furthermore, we grouped 5 chief complaints with very low occurrences, such as "drowning accident" or "flu vaccine," into our class "diverse." However, we did not use this group in further analysis because of the heterogeneity of the symptoms included. The lead symptom topics were then aggregated into 12 clusters by the authors according to clinical judgment. The complete list of LS can be found in Table S1 in Multimedia Appendix 1.

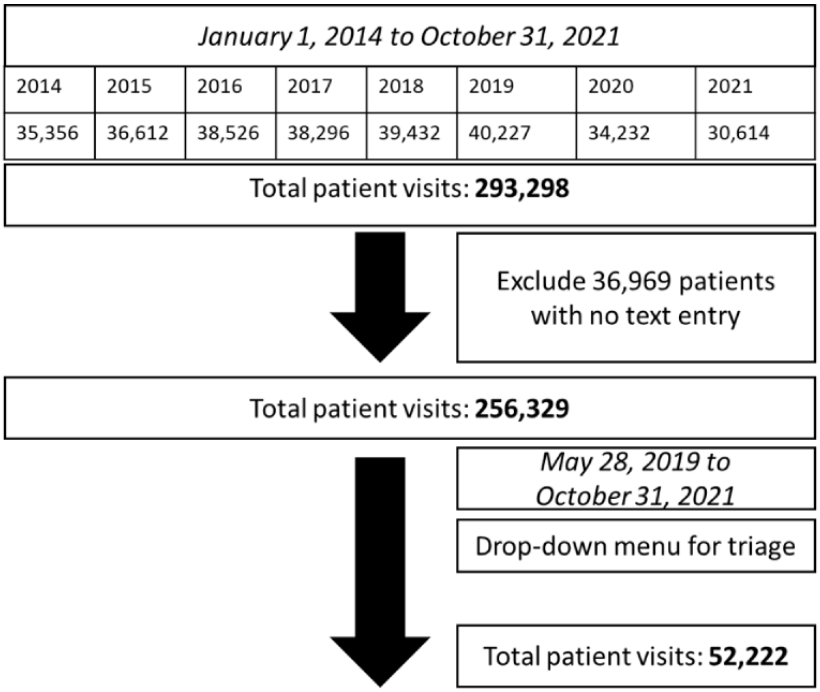
A total of 65 variables from 2014 to 2018 and 69 variables from 2019 to 2021 (including the chief complaint) were recorded in the initial data set. A total of 65 variables from 2014 to 2018 were constant throughout 2014 to 2021 and were retained for preprocessing. The final data table used for the analysis contained the variables listed in Table 1, in addition to the patient ID, year and weekday of the consultation derived from the admission time stamp, the treating ED team (internal medicine, surgery, neurology, or psychiatry), as well as the LS clusters from the drop-down menu and the NLP-extracted topic clusters that were obtained from the field "suspected diagnoses," discussed in detail in *Analysis: Topic Allocation* section. In addition, the table contained the outcomes "inpatient" and "ESI score<4" as binary variables. Two further outcomes were considered, namely, readmission within 30 days and waiting time>30 minutes, but were discarded owing to doubts regarding the quality and consistency of the entered data. We retained the outcomes "inpatient" and "ESI score<4" owing to their direct association with the immediacy of the outcome in the patient pathway within the ED, ensuring robust data quality.

Table 1. Summary statistics of the patient population (n=52,222)[a].

Variable	Values
Age (y), mean (SD)	46.5 (19.7)
Female, n (%)	23,782 (45.54)
Emergency Severity Index score (out of 5), mean (SD)	3.3 (0.6)
Fast track, n (%)	8264 (15.82)
Number of patients in the emergency department, mean (SD)	19.8 (8.3)
Early shift, n (%)	21,644 (41.45)
With emergency medical service, n (%)	9020 (17.27)
With police, n (%)	188 (0.36)
Accident, n (%)	16,845 (32.26)
Inpatient, n (%)	14,112 (27.02)
Night shift, n (%)	7915 (15.16)
Late shift, n (%)	22,663 (43.4)

[a]The total sample contains patient visits for the period from May 28, 2019, to October 31, 2021.

Figure 2. Data structure.



Textbox 1. Variables recorded for our analysis.

Triage

- Suspected diagnosis (free text) and Emergency Severity Index score

Type of discharge

- Hospitalization, ambulatory treatment, or patient has run away

Patient characteristics

- Patient visit pseudo ID, age, gender, admission type (self, ambulance, or police), and admission reason (accident or illness)

Organizational

- Average number of patients in emergency department (ED); average patient waiting time; night, late, or early shift; and treating ED team (internal medicine, surgery, neurology, or psychiatry)

Time

- Time stamp

Analysis: Topic Allocation

We selected the field "suspected diagnosis" to extract the symptoms or complaints that led to ED admission according to the oral report received by the ED physician in charge, as mentioned previously. This field comprises a short-text string entered by the ED physician upon receiving information about the patient's expected arrival at the ED. This information can be transmitted to the ED physician by a referring physician or ambulance well in advance of a patient's arrival. The text is entered before the patient triage is performed by the triage ED nurse. As a clinical note, the physician's text entry is part of the EHR. The information contained in the string "suspected diagnoses" is supposed to be similar to the selected chief complaint from the drop-down menu "lead symptom." Indeed, the latter variable was added later (in 2019) to facilitate the administrative analysis of causes for ED admission, as an analysis using unstructured text was not possible by the hospital administration. Both fields are supposed to contain the medical reason, or chief complaint, leading to ED admission.

We constructed a measure of the semantic distance of all words in the corpus by training a word embedding. Word embeddings are matrices in which each column represents a word and its relative distance to other words (eg, the distance between blood and red is smaller than that between blood and green). Hence, it is possible to find the most similar words for a given keyword using the smallest distance measured with the cosine similarity. To train the word embedding, we used word2vec with the entire text corpus and the continuous bag-of-words algorithm from the Python library Gensim [16], with an embedding size of 300 computed with 100 epochs.

To construct our topic dictionaries, we proceeded in 4 steps, as shown in Figure 3. First, we manually defined topics and selected between 2 and 20 initial seed words (henceforth "keywords") by reading some of the texts and using prior medical knowledge. A smaller number of keywords were used for the design of the topic "infection" (n=1). A larger number of initial keywords were used for the design of the topics "intoxication" (n=40) and "skin" (n=28). In step 2, we then searched for up to 50 of the semantically closest words for each initial list. With the help of the word embedding, it is possible to search for the words that maximize the cosine similarity for the seed keywords. In addition, we only considered keywords that occurred at least 10 times. This list of similar words allowed us to efficiently increase the dictionary for each topic. In step 3, we manually chose words from the preselection of similar words to the seed word, resulting in a separate dictionary per topic (step 4). In some instances, the dictionary used combinations of words. For instance, the topic "chest pain" was allocated to combinations of words such as "pain" or "pressure" with the words "chest" or "thorax."

This table presents the distribution of the diagnosis topics obtained with the NLP-based text annotation before and after the spherical feature annotation. The total number of cases was 52,222, and 20.38% could not be attributed with a diagnosis topic.

The summary of the increase in tags per topic cluster through the NLP-based expansion of our topics library is presented in Table 2. The first column shows the percentage of the sample tagged with a topic using the original keyword approach. The proportion of clinical topics ranged from 0.72% for COVID-19 to 31.6% for trauma-related visits. It should be noted that patient visits can be allocated with multiple topics. The next column shows the share of visits with the spherically increased dictionary, with the percentage increase in topic shares in the last column. Overall, the spherical dictionary enhancement decreased the number of nontagged visits by nearly 25%, from 27.08% of the sample to 20.24%. For the individual topics, the additional keywords increased their share, ranging from 5.29% for trauma to 286.35% for general administrative visits.

In the second procedure, we automatically increased the number of keywords for each topic dictionary. This process is shown in Figure 4, which can be imagined as constructing a multidimensional sphere using the initial keywords. The additional keywords were then located within that sphere.

The "spherical" dictionary enhancement consists of the following steps:

- Compute all distances between the keywords and retain the largest distance (ie, the distance between the 2 least similar words). For each keyword, this distance is the radius of a circle in the embedding space (steps 1 and 2).
- For each of the initial keywords, identify the n-closest words (not in the topic dictionary) using the cosine similarity (step 3).
- Retain these additional words if their distance to all other initial keywords is smaller than the maximum distance computed in the first step, that is, if the new words are in the intersection of all circles (step 4).

Using the abovementioned approach, we could tag 79.76% (41,653/52,222) of the final sample. The remaining texts could not be tagged because they either belonged to small topics that we did not define or because these texts did not contain words that are present in the dictionary.

Once the dictionaries for each topic are constructed, they can be used for additional patient visits and for similar data sets, which makes the approach easily scalable.

Figure 3. Topic dictionaries with semimanual keyword selection. (A) The researcher selects an initial seed word for a topic. (B) Using word embeddings, a list of semantically similar words from the corpus is generated. (C) The researcher manually selects words that are associated with the topic. (D) The topic dictionary is created.

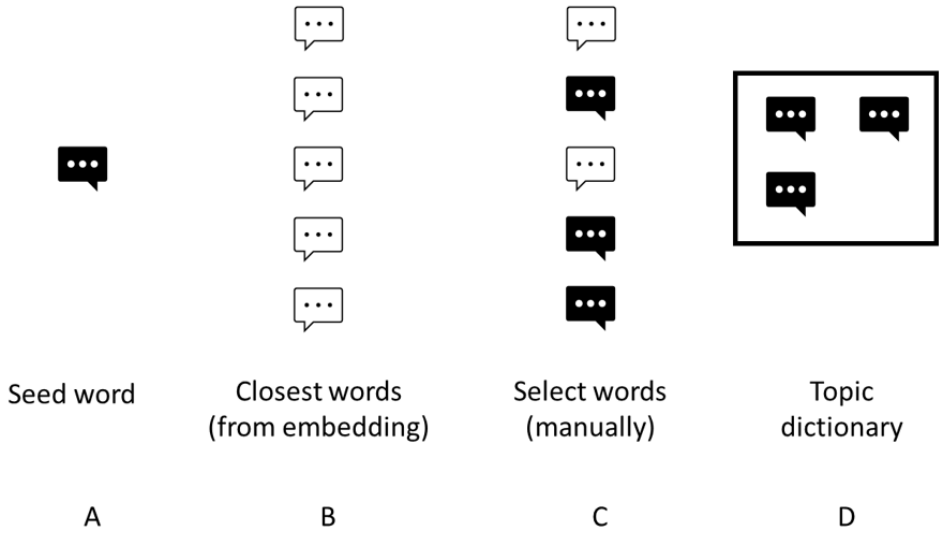


Table 2. Spherical feature annotation and increase in topic share (n=52,222)[a].

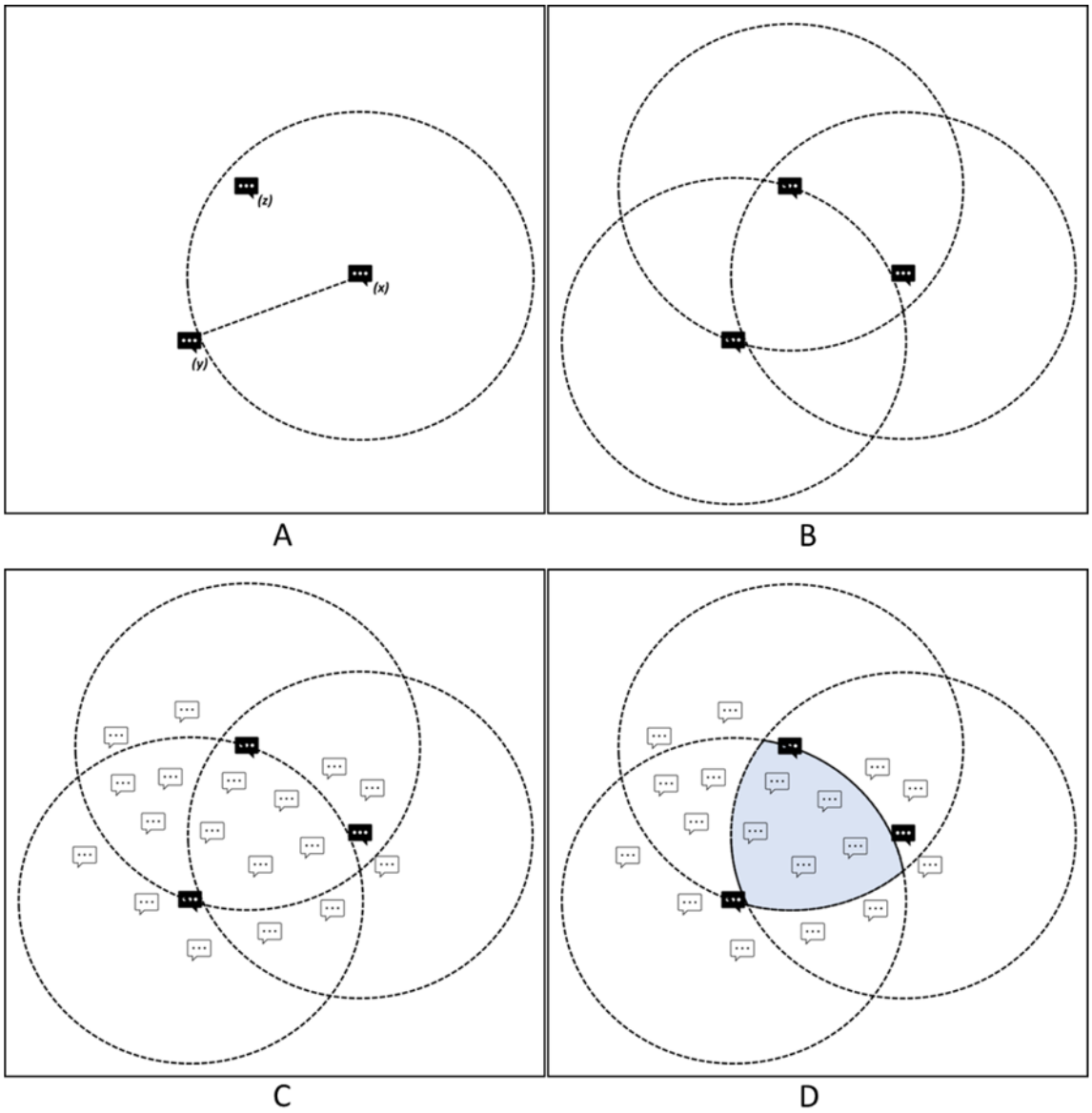
Clinical topic NLP[b]	Records tagged initially, n (%)	Records tagged NLP-augmented, n (%)	Increase in tagged patient records, n (%)[c]
COVID-19	375 (0.72)	405 (0.78)	30 (8)
General symptom	6401 (12.26)	6867 (13.15)	466 (7.28)
General administration	315 (0.6)	1217 (2.33)	902 (286.35)
Systemic clinical	3219 (6.16)	3519 (6.74)	300 (9.32)
Gastrointestinal	3421 (6.55)	4159 (7.96)	738 (21.57)
Respiratory	4040 (6.55)	4159 (7.96)	738 (21.57)
Cardiovascular	2683 (5.14)	5219 (9.99)	2536 (94.52)
Neurological	414 (7.93)	4485 (8.59)	345 (8.33)
Eye; ear, nose, and throat; and derma	1818 (3.48)	2061 (3.95)	243 (13.37)
Gynecology and urology	2712 (5.19)	3004 (5.75)	292 (10.77)
Trauma	16,516 (31.63)	17,389 (33.3)	873 (5.29)
General psychiatric	1989 (3.81)	2627 (5.03)	638 (32.08)
No tag	14,141 (27.08)	10,569 (20.24)	–3572 (–25.26)

[a]This table presents the distribution of the diagnosis topics obtained with the NLP-based text annotation before and after the spherical feature annotation.

[b]NLP: natural language processing.

[c]Percent of initially recorded tags.

Figure 4. Spherical dictionary enhancement. (A) Step A uses the largest distance between 2 words that are already in the topic. The circle around the word (x) shows the region in the embedding space with words closer to (x) than the maximum distance. (B) The same region is circled around the other 2 words (y) and (z). (C) The other words in the embedding space that were initially not included in the topic. (D) The intersection of the 3 circles defines the area in the embedding space where the distance of each word is smaller than the maximum distance.



Ethical Considerations

A waiver from the cantonal ethics committee was obtained before the commencement of this study (BASEC-Nr. Req-2019-00671).

Results

In the first step, we performed a descriptive analysis of the topics. To this end, we first excluded cases without a manually selected LS for further analysis and obtained a data set with 52,222 entries. Of the 52,222 patient visits included in our final analysis, 5994 (11.48%) had a manually recorded chief complaint that was not otherwise specified (eg, “Diverse”) and could not be classified as a symptom Of the 52,222 entries, 10,569 (20.24%) were not tagged with an NLP topic.

The distribution of all NLP topics is shown in Table 3. The distribution ranged from 0.05% of patient visits tagged with the NLP topic “dementia” to 9.89% for “wound.” The largest cluster of aggregated NLP symptom-related topics was “trauma,” with 33.1% of visits, and the smallest was “COVID,” with 0.8% of visits. The distribution of chief complaints can be found in Table S1 in Multimedia Appendix 1. In total, the distribution ranged from 0.01% of patient visits for the recorded chief complaints “melaena,” “hearing problems,” and “contact with chemicals” to 14.6% for “COVID.” The largest cluster of aggregated chief complaints was “trauma” with 23.6% and the smallest was “general organizational” with 1.2% of visits.

For comparability, we grouped all LS and NLP topics into 12 identical symptom clusters, which can be found in Table 4.

Table 3. Clusters for natural language processing–extracted topics (n=52,222)[a].

Cluster and subcluster detail	Values, n (%)
COVID-19	401 (0.77)
General symptoms	6852 (13.12)
Fever	2440 (4.67)
Pain	4505 (8.63)
General weakness	80 (0.15)
Back pain	438 (0.84)
General organizational	1217 (2.33)
Follow-up and prescription	1217 (2.33)
Systemic	3519 (6.74)
Infection not otherwise specified	1239 (2.37)
Sepsis	125 (0.24)
Anaphylaxia and allergy	261 (0.5)
Cancer	1688 (3.23)
Transplantation	227 (0.43)
Glycemia	138 (0.26)
Gastrointestinal	4147 (7.94)
Gastrointestinal bleeding	522 (1)
Abdominal pain	1879 (3.6)
Diarrhea, vomiting, and nausea	2248 (4.3)
Respiratory	4311 (8.26)
Upper airway	1592 (3.05)
Lower airway	1934 (3.7)
Influenza	440 (0.84)
Dyspnea	2197 (4.21)
Cardiovascular	5211 (9.98)
Chest pain	3569 (6.83)
Palpitations and arrythmia	518 (0.99)
Pulmonary embolism	281 (0.54)
Deep venous thrombosis	528 (1.01)
Hypertension	394 (0.75)
Neurological	4466 (8.55)
Headache	1189 (2.28)
Neurological	1737 (3.33)
Vigilance and disorientation	191 (0.37)
Dementia	24 (0.05)
Syncope	453 (0.87)
Vertigo and dizziness	934 (1.79)
Convulsion	226 (0.43)
Eye; ear, nose, and throat; and skin	2061 (3.95)
Epistaxis	58 (0.11)
Eye symptoms	703 (1.35)
Hearing and auricular	18 (0.03)

Cluster and subcluster detail	Values, n (%)
Skin	1311 (2.51)
Urological and gynecological	3004 (5.75)
Urological and kidney	2973 (5.69)
Pregnancy	34 (0.07)
Trauma	17,302 (33.13)
Wound	5163 (9.89)
Fracture and luxation	5375 (10.29)
Trauma and head	2171 (4.16)
Burns	141 (0.27)
Fall	729 (1.4)
Trauma not otherwise specified	9278 (17.77)
Bleeding not otherwise specified	986 (1.89)
Collision	1250 (2.39)
Traffic	314 (0.6)
Psychiatric	2625 (5.03)
Intoxication	1146 (2.19)
Psychiatric	851 (1.63)
Fear	725 (1.39)
Severity	
Nonsevere	113 (0.22)
Severe	235 (0.45)
Chronic	55 (0.11)
Acute	232 (0.44)

[a]This table presents the distribution of the diagnosis topics obtained with the natural language processing–based text annotation. In total, 20.38% of cases could not be attributed with a diagnosis topic.

Table 4. Summary statistics feature annotations (n=52,222)[a].

Cluster	LS[b], (n)	NLP[c] (n)	Correlation (*r*)[d]	Consistency[e]
COVID-19	7623	401	0.18	0.05
General symptom	7993	6852	−0.04	0.10
General administration	642	1217	0.01	0.04
Systemic clinical	1983	3519	0.12	0.22
Gastrointestinal	4063	4147	0.41	0.46
Respiratory	872	4311	0.17	0.44
Cardiovascular	2245	5211	0.28	0.49
Neurological	5123	4466	0.44	0.46
Eye; ear, nose, and throat; and derma	1041	2061	0.26	0.39
Gynecology and urology	1206	3004	0.40	0.67
Trauma	12,337	17,302	0.54	0.79
General psychiatric	1610	2625	0.60	0.78
No tag	5994	10,644	0.07	0.28

[a]This table presents the number of tagged cases for each chief cluster with both the natural language processing–based method and based on the chief complaint tag.

[b]LS: lead symptom.

[c]NLP: natural language processing.

[d]Correlation between LS and NLP.

[e]The number of overlapping LS and NLP tags divided by the total number of LS tags.

In addition to the NLP symptom-related topics, 4 modulating NLP topics, “acute,” “chronic,” “nonsevere,” and “severe,” were recorded, also based on keywords (ie, words in the text indicating severity). The purpose of the modulating topics is to provide more information on severity and control for this dimension in the further analysis.

We found that the correlation between LS clusters and NLP clusters was low (Table 4). Similarly, consistency varies relative to the LS. We also calculated the consistency of the NLP tags relative to the LS groups (the LS groups are the denominator; being more established, we use them as a benchmark). For most clusters, the consistency is approximately 50%, with trauma and psychiatric diagnosis having the highest consistency of 78% and 79%, respectively, and general administration and COVID-19 having the lowest consistency of 4% and 5%, respectively.

Compared with the LS clusters, our NLP topics have the advantage that a patient visit can be tagged to multiple topics. Table S2 in Multimedia Appendix 1 shows the number of NLP topics for each LS cluster. Of the 46,228 patient visits where we could assign a manually recorded chief complaint, 8950 (19.36%) were not tagged with an NLP topic. In contrast, 33.48% (15,477/46,228) of the visits were tagged with at least 2 NLP topics.

We estimated 3 models using logistic regression to show the association of the different symptom groups with the ESI and inpatient indicators:

$$\text{Model 1: } Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i \textbf{ (1)}$$

$$\text{Model 2: } Y_i = \alpha + \beta X_i + \delta W_i + \varepsilon_i \textbf{ (2)}$$

$$\text{Model 3: } Y_i = \alpha + \beta X_i + \gamma Z_i + \delta W_i + \varepsilon_i \textbf{ (3)}$$

where Y_i is either the ESI or inpatient indicator variable for patient visit i, α the intercept, X_i is a vector of demographic and organizational variables for patient visit I (age; gender; admission type; admission reason; average number of patients in ED; average patient waiting time; night, late, or early shift; and treating ED team), Z_i is a vector of the NLP-derived symptom clusters, W_i is a vector of the lead symptom–derived cluster (based on the drop-down menu), and ε_i is the error term.

Tables 5 and 6 present the results. Column 1 shows the NLP-derived groups, with coefficients ranging between 5% and 13% increased or decreased probability of a high ESI score or 5% to 19% increased or decreased probability for hospitalization. The drop-down–based LS in column 2 has similar but slightly larger coefficients. Column 3 shows both variables, as in model 3, in this specification, the coefficients are mostly complementary, meaning that if a patient shows the same symptom in both the NLP and LS measures, the probabilities can be added. Note that this is not owing to multicollinearity because both coefficients remain significant in most cases.

Table 5. Linear probability model on "Inpatient"[a].

Name of cluster[b]	Model 1[c], regression coefficient (SE)	Model 2[c], regression coefficient (SE)	Model 3 including both measures[d], regression coefficient (SE)
NLP[e] cluster: COVID-19	0.048[f] (0.019)	N/A[g]	−0.022 (0.022)
Chief complaint cluster: COVID-19	N/A	0.127[g] (0.007)	0.133[h] (0.008)
NLP cluster: general symptoms	0.011[f] (0.005)	N/A	−0.019[h] (0.005)
Chief complaint cluster: general symptoms	N/A	−0.002 (0.007)	0.000 (0.007)
NLP cluster: general organizational	−0.004 (0.011)	N/A	0.006 (0.011)
Chief complaint cluster: general organizational	N/A	−0.062[g] (0.016)	−0.052[h] (0.016)
NLP cluster: systemic	0.117[h] (0.007)	N/A	0.101[h] (0.007)
Chief complaint cluster: systemic	N/A	0.118[h] (0.010)	0.104[h] (0.010)
NLP cluster: gastrointestinal	0.071[h] (0.006)	N/A	0.040[h] (0.007)
Chief complaint cluster: gastrointestinal	N/A	0.083[h] (0.008)	0.059[h] (0.008)
NLP cluster: respiratory	0.063[h] (0.007)	N/A	−0.017[f] (0.008)
Chief complaint cluster: respiratory	N/A	0.133[h] (0.014)	0.126[f] (0.014)
NLP cluster: cardiovascular	−0.020[h] (0.006)	N/A	−0.009 (0.006)
Chief complaint cluster: cardiovascular	N/A	−0.038[h] (0.010)	−0.031[h] (0.010)
NLP cluster: neurological	−0.046[h] (0.007)	N/A	−0.045[h] (0.007)
Chief complaint cluster: neurological	N/A	−0.058[h] (0.009)	−0.048[h] (0.009)
NLP cluster: eye, ENT[i], or skin	−0.055[h] (0.009)	N/A	−0.044[h] (0.009)
Chief complaint cluster: eye, ENT, or skin	N/A	−0.128[h] (0.013)	−0.112[h] (0.013)
NLP cluster: urological or gynecological	−0.015[f] (0.008)	N/A	−0.004 (0.008)
Chief complaint cluster: urological or gynecological	N/A	−0.033[h] (0.012)	−0.036[h] (0.013)
NLP cluster: trauma	−0.041[h] (0.005)	N/A	−0.038[h] (0.005)
Chief complaint cluster: trauma	N/A	0.011 (0.007)	0.020[h] (0.007)
NLP cluster: psychiatric	−0.079[h] (0.009)	N/A	−0.053[h] (0.010)
Chief complaint cluster: psychiatric	N/A	−0.068[h] (0.013)	−0.039[h] (0.014)

[a]This table presents the results from a linear probability model with inpatients as the dependent variable. All the models include a set of demographic and administrative covariates.

[b]Observation: 52,222; R^2=0.259.

[c]Observation: 52,222; R^2=0.263.

[d]Observation: 52,222; R^2=0.269.

[e]NLP: natural language processing.

[f]P<.05.

[g]N/A: not applicable.

[h]P<.01.

[i]ENT: ear, nose, and throat.

Table 6. Linear probability model on "low Emergency Severity Index (ESI) score"[a].

Name of cluster[b]	Model 1[c], regression coefficient (SE)	Model 2[c], regression coefficient (SE)	Model 3 including both measures[d], regression coefficient (SE)
NLP[e] cluster: COVID-19	0.079[f] (0.019)	N/A[g]	0.023 (0.019)
Chief complaint cluster: COVID-19	N/A	0.214[f] (0.007)	0.172[f] (0.007)
NLP cluster: general symptoms	0.036[f] (0.005)	N/A	−0.023[f] (0.005)
Chief complaint cluster: general symptoms	N/A	−0.142[f] (0.007)	0.127[f] (0.007)
NLP cluster: general organizational	−0.050 (0.011)	N/A	−0.044[f] (0.011)
Chief complaint cluster: general organizational	N/A	0.308[f] (0.016)	0.352[f] (0.016)
NLP cluster: systemic	0.076[f] (0.007)	N/A	0.093[f] (0.007)
Chief complaint cluster: systemic	N/A	0.009 (0.010)	0.009 (0.010)
NLP cluster: gastrointestinal	0.192[f] (0.006)	N/A	0.088[f] (0.007)
Chief complaint cluster: gastrointestinal	N/A	0.305[f] (0.008)	0.262[f] (0.008)
NLP cluster: respiratory	0.114[f] (0.007)	N/A	0.053[f] (0.007)
Chief complaint cluster: respiratory	N/A	0.121[f] (0.014)	0.088[f] (0.014)
NLP cluster: cardiovascular	0.050[f] (0.006)	N/A	0.030[f] (0.006)
Chief complaint cluster: cardiovascular	N/A	0.205[f] (0.009)	0.197[f] (0.010)
NLP cluster: neurological	−0.015[h] (0.007)	N/A	−0.002 (0.007)
Chief complaint cluster: neurological	N/A	−0.038[f] (0.009)	−0.039[f] (0.009)
NLP cluster: eye, ENT[i], or skin	−0.134[f] (0.009)	N/A	−0.061[f] (0.009)
Chief complaint cluster: eye, ENT, or skin	N/A	−0.302[f] (0.013)	−0.279[f] (0.013)
NLP cluster: urological or gynecological	0.055[f] (0.008)	N/A	0.006 (0.008)
Chief complaint cluster: urological or gynecological	N/A	0.193[f] (0.012)	0.187[f] (0.013)
NLP cluster: trauma	−0.129[f] (0.005)	N/A	−0.098[f] (0.005)
Chief complaint cluster: trauma	N/A	−0.011 (0.007)	0.013[j] (0.007)
NLP cluster: psychiatric	0.063[f] (0.009)	N/A	0.080[f] (0.010)
Chief complaint cluster: psychiatric	N/A	0.086[f] (0.012)	0.051[f] (0.013)

[a]This table presents the results from a linear probability model with the low ESI score indicator as the dependent variable (ESI score of 2 or 3). All models included a set of demographic and administrative covariates.

[b]Observation: 52,222; R^2=0.409.

[c]Observation: 52,222; R^2=0.448.

[d]Observation: 52,222; R^2=0.457.

[e]NLP: natural language processing.

[f]P<.01.

[g]N/A: not applicable.

[h]P<.05.

[i]ENT: ear, nose, and throat.

[j]P<.10.

Of the 12 symptom clusters, 11 (92%) in column 1 had a significant regression coefficient for hospitalization (all but "general organizational"). Eight clusters remained significant even when including the cluster of clinician-determined chief complaints in the model. In the model explaining "inpatient," in 10 (83%) out of the 12 symptom cluster pairs, the coefficients of the NLP topic clusters showed the same algebraic sign as the chief complaint clusters. In contrast, for 2 symptom cluster pairs, they did not ("general symptoms" and "trauma"). A change in the algebraic sign of either the chief complaint cluster

or the NLP topics cluster occurred in 4 cluster pairs when both NLP topics and chief complaints were included in the model ("COVID," "general symptoms," "general organizational," and "respiratory"). We obtained similar results when analyzing the low ESI scores. However, a change in the algebraic sign of a coefficient within solely 1 pair of symptom clusters was noted ("trauma"). Interestingly, the clusters "cardiovascular," "neurological," and "trauma" were significantly associated with nonhospitalization, of which "neurological" and "trauma" but not "cardiovascular" were also significantly associated with a lower ESI score.

As a robustness check, we used each of the 3 model specifications to predict the ESI indicator and the inpatient indicator. Using the respective sets of variables of each specification, we used a logistic regression with a 2:1 train-test split to predict both outcomes. Table 7 shows the F_1-score and area under the curve (AUC) score of these predictions. The results show that the 3 specifications have similar predictive power (an AUC of 0.82-0.84 for "inpatient" and an AUC of 0.90-0.92 for ESI indicator).

The inference and prediction results show that the added value of text in this setting is not by increasing the predictive power of the model, where the outcomes are existing process outcomes (eg, discharge type of severity). Instead, unstructured text allows clinicians to access more granular information to optimize patient flows, which cannot be reflected in the inpatient and ESI indicator outcomes.

In a more granular analysis, we estimated models 1 to 3 with the individual NLP topics and the individual LS groups instead of the clusters previously used. The analysis corroborated our clinical presumptions that, for example, age, admission by an ambulance, and "sepsis" as an NLP topic, as well as "chest pain" for a chief complaint, were associated with low ESI scores (2 or 3) or hospital admission. In contrast, the NLP topic or chief complaint cluster "follow-up" was not. The complete results are provided in Tables S3-S6 in Multimedia Appendix 1.

Table 7. Prediction of hospitalization ("Inpatient") and low Emergency Severity Index (ESI) score of 2 or 3 ("Low ESI score").

Variable and model	F_1-score on ones	AUC[a]
Inpatient		
Model 1: NLP[b] clusters	0.57	0.82
Model 2: LS[c] clusters	0.57	0.83
Model 3: NLP+LS clusters	0.59	0.84
Low ESI score		
Model 1: NLP clusters	0.86	0.92
Model 2: LS clusters	0.84	0.90
Model 3: NLP+LS clusters	0.87	0.92

[a]AUC: area under the curve.

[b]NLP: natural language processing.

[c]LS: lead symptom.

Discussion

Principal Findings

Our analysis of patient records showed the additional information extracted from unstructured text and its potential usefulness in the clinical context. We demonstrated that the information extracted from NLP features and the physician's categorization of chief complaints was *complementary*. Indeed, the correlation and consistency between the chief complaint and NLP-derived clusters were low (Table 4). This finding indicates that the free text from the NLP clusters provides additional information than that contained in the symptom clusters from the structured chief complaints.

The complementarity of the information is further emphasized by the results summarized in Tables 5 and 6, and most coefficients remained significant when both types of indicators were included in the model, suggesting that different aspects of patient information appear to be encoded by the 2 approaches. These results support our hypothesis that NLP-derived libraries capture greater depth and breadth of information than a single chief complaint and underscore the relevance of including information captured in unstructured text to address patient populations.

Surprisingly, the "cardiovascular" and "trauma" clusters were not significant features for predicting hospitalization, with "trauma" also significant for predicting a *higher* ESI score. In contrast, the "systemic" cluster, which included sepsis, anaphylaxis, and neoplastic disease, was significant for predicting hospitalization and a lower ESI score, consistent with clinical expectations. Although symptoms suggestive of cardiac dysfunction and trauma may warrant urgent clinical risk assessment, most patients with such complaints would not require hospitalization. Therefore, early allocation of hospital beds for these subgroups is unlikely to reduce overcrowding. Targeting patients with systemic symptoms, in contrast, is likely to do so.

We also proposed a method for analyzing unstructured clinical notes. Our approach has the advantages of speed, simplicity of

implementation, and transparency. The speed at which supervised libraries can be assembled is a strength of the proposed approach. A limitation of implementing supervised NLP algorithms in routine decision support is that they are often resource intensive [17]. In our application, it took an untrained clinician only a few days to assemble the entire library.

Furthermore, using NLP as a tool traditionally requires expertise and the ability to master NLP applications. In fields that require years to decades of training, such as health care, professionals cannot be routinely trained to excel in programming. Thus, a further major barrier to the successful implementation of NLP applications in health care is often the usability of NLP applications [18]. Moreover, the flexibility of the method allows easy adaptation of the created dictionaries to analyze new data sets.

Trust is one of the key benefits of clinician involvement in developing proprietary AI models. Indeed, lack of trust is a recognized major limitation that hinders the potential benefits of using AI in routine clinical practice for organizations and patients [19,20]. Owing to the supervised approach, annotated library compilation is comprehensible and transparent; hence, it is trustworthy for clinicians. This may also become an important advantage if regulation on the implementation of AI use in health care tightens in the future.

The limitation of this study is that our approach still requires manual coding. However, future developments in AI may facilitate this step even further. In addition, human bias was possible because the library was compiled manually. In general, an AI-based text analysis does not achieve perfect precision. However, we primarily advocate using free-text analysis for organizational, not clinical, decision support. Therefore, this limitation is not clinically relevant. A further limitation may lie in the fact that the low correlation between the NLP and chief complaint clusters could stem from errors originating from the manual grouping or NLP clustering. However, we believe these results are plausible. Indeed, the chief complaints “fever” and “pain” were included in the cluster “general symptoms,” as were the NLP-extracted tags “fever” and “pain.” However, as only 1 chief complaint could allocated to a patient, during the COVID-19 pandemic, most patients presenting with fever or influenza-like pain would have most likely been categorized as presenting with the chief complaint “COVID.”

Conclusions

Health care workers on the one side and EHR engineers as well as hospital administration on the other side are caught in a long, ongoing conflict over the extent of structuring the data entered into EHR. Health care workers often argue that entering structured data is a cumbersome task and that the information archived can be of little use in daily clinical practice. In contrast, administrators and EHR engineers often advocate that structuring data is the only reliable solution, enabling a meaningful analysis of the data. Technological advances may help resolve this conflict.

We were able to demonstrate the importance of maintaining free text in EHR. Indeed, using the chief complaints attributed by a physician from a drop-down menu and a corresponding free-text field as a case in point, we were able to show that free text contains a wealth of information that is not routinely captured by structured data.

Moreover, we developed an approach that could enable the information captured in free text to be easily extracted and processed by hospital informatics systems and fed into a workflow, possibly improving the efficiency of patient management.

Therefore, future EHRs should include the possibility of entering free text.

Acknowledgments

The authors would like to thank Professor Michael Krauthammer from the University of Zurich, Switzerland, and Privat-Dozentin Dr Ksenija Slankamenac, PhD, from the University Hospital Zurich for their feedback in helping to prepare this submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary file for web-based publication only.
[PDF File (Adobe PDF File), 240 KB - medinform_v12i1e49007_app1.pdf]

References

1. Hwang JE, Seoung BO, Lee SO, Shin SY. Implementing structured clinical templates at a single tertiary hospital: survey study. JMIR Med Inform 2020 Apr 30;8(4):e13836. [doi: 10.2196/13836] [Medline: 32352392]
2. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18(2):181-186 [FREE Full text] [doi: 10.1136/jamia.2010.007237] [Medline: 21233086]
3. 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]

4. 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. [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
5. Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. J Am Med Inform Assoc 2022 Jun 14;29(7):1292-1302 [FREE Full text] [doi: 10.1093/jamia/ocac058] [Medline: 35475536]
6. Velt KB, Cnossen M, Rood PP, Steyerberg EW, Polinder S, Lingsma HF. Emergency department overcrowding: a survey among European neurotrauma centres. Emerg Med J 2018 Jul;35(7):447-448 [FREE Full text] [doi: 10.1136/emermed-2017-206796] [Medline: 29563151]
7. Di Somma S, Paladino L, Vaughan L, Lalle I, Magrini L, Magnanti M. Overcrowding in emergency department: an international issue. Intern Emerg Med 2015 Mar;10(2):171-175. [doi: 10.1007/s11739-014-1154-8] [Medline: 25446540]
8. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: a systematic review of causes, consequences and solutions. PLoS One 2018 Aug 30;13(8):e0203316 [FREE Full text] [doi: 10.1371/journal.pone.0203316] [Medline: 30161242]
9. Iacobucci G. Overcrowding and long delays in A&E caused over 4000 deaths last year in England, analysis shows. BMJ 2021 Nov 18;375:n2835. [doi: 10.1136/bmj.n2835] [Medline: 34794954]
10. Iacobucci G. Government must "get a grip" on NHS crisis to halt avoidable deaths, say leaders. BMJ 2023 Jan 03;380:12. [doi: 10.1136/bmj.p12] [Medline: 36596573]
11. Boyle A. Unprecedented? The NHS crisis in emergency care was entirely predictable. BMJ 2023 Jan 09;380:46. [doi: 10.1136/bmj.p46] [Medline: 36623878]
12. Boonstra A, Broekhuis M. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. BMC Health Serv Res 2010 Aug 06;10:231 [FREE Full text] [doi: 10.1186/1472-6963-10-231] [Medline: 20691097]
13. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022.
14. Tang J, Meng Z, Nguyen X, Mei Q, Zhang M. Understanding the limiting factors of topic modeling via posterior contraction analysis. In: Proceedings of the 31st International Conference on International Conference on Machine Learning ICML'14. 2014 Presented at: ICML'14; June 21-26, 2014; Beijing, China.
15. Maynard D, Funk A. Automatic detection of political opinions in Tweets. In: Proceedings of the Workshops at the 8th Extended Semantic Web Conference, ESWC 2011. 2011 Presented at: Workshops at the 8th Extended Semantic Web Conference, ESWC 2011; May 29-30, 2011; Heraklion, Greece. [doi: 10.1007/978-3-642-25953-1_8]
16. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010 Presented at: LREC 2010 Workshop on New Challenges for NLP Frameworks; May 22, 2010; Valletta, Malta. [doi: 10.13140/2.1.2393.1847]
17. 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 Dec 17;2(1):130 [FREE Full text] [doi: 10.1038/s41746-019-0208-8] [Medline: 31872069]
18. Zheng K, Vydiswaran VG, Liu Y, Wang Y, Stubbs A, Uzuner Ö, et al. Ease of adoption of clinical natural language processing software: an evaluation of five systems. J Biomed Inform 2015 Dec;58 Suppl(Suppl):S189-S196 [FREE Full text] [doi: 10.1016/j.jbi.2015.07.008] [Medline: 26210361]
19. 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]
20. Celi LA, Fine B, Stone DJ. An awakening in medicine: the partnership of humanity and intelligent machines. Lancet Digit Health 2019 Oct;1(6):e255-e257 [FREE Full text] [doi: 10.1016/s2589-7500(19)30127-x] [Medline: 32617524]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
ED: emergency department
EHR: electronic health record
ESI: Emergency Severity Index
LS: lead symptoms
NLP: natural language processing

Edited by C Lovis; submitted 15.05.23; peer-reviewed by J Kors, C Gaudet-Blavignac; comments to author 30.06.23; revised version received 30.10.23; accepted 24.11.23; published 17.01.24.

Please cite as:
Mehra T, Wekhof T, Keller DI
Additional Value From Free-Text Diagnoses in Electronic Health Records: Hybrid Dictionary and Machine Learning Classification Study
JMIR Med Inform 2024;12:e49007
URL: https://medinform.jmir.org/2024/1/e49007
doi:10.2196/49007
PMID:38231569

Original Paper

BERT-Based Neural Network for Inpatient Fall Detection From Electronic Medical Records: Retrospective Cohort Study

Cheligeer Cheligeer[1,2], PhD; Guosong Wu[1,3], PhD; Seungwon Lee[1,2], PhD; Jie Pan[1,3], PhD; Danielle A Southern[1], MSc; Elliot A Martin[1,2], PhD; Natalie Sapiro[1], MSc, RN; Cathy A Eastwood[1,3], RN, PhD; Hude Quan[1,3], MD, PhD; Yuan Xu[1,3,4,5], MD, PhD

[1]Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[2]Provincial Research Data Services, Alberta Health Services, Calgary, AB, Canada

[3]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[4]Department of Oncology, University of Calgary, Calgary, AB, Canada

[5]Department of Surgery, University of Calgary, Calgary, AB, Canada

Corresponding Author:
Yuan Xu, MD, PhD
Centre for Health Informatics, Cumming School of Medicine
University of Calgary
3280 Hospital Dr NW
Calgary, AB, T2N 4Z6
Canada
Phone: 1 (403) 210 9554
Email: yuxu@ucalgary.ca

Abstract

Background: Inpatient falls are a substantial concern for health care providers and are associated with negative outcomes for patients. Automated detection of falls using machine learning (ML) algorithms may aid in improving patient safety and reducing the occurrence of falls.

Objective: This study aims to develop and evaluate an ML algorithm for inpatient fall detection using multidisciplinary progress record notes and a pretrained Bidirectional Encoder Representation from Transformers (BERT) language model.

Methods: A cohort of 4323 adult patients admitted to 3 acute care hospitals in Calgary, Alberta, Canada from 2016 to 2021 were randomly sampled. Trained reviewers determined falls from patient charts, which were linked to electronic medical records and administrative data. The BERT-based language model was pretrained on clinical notes, and a fall detection algorithm was developed based on a neural network binary classification architecture.

Results: To address various use scenarios, we developed 3 different Alberta hospital notes-specific BERT models: a high sensitivity model (sensitivity 97.7, IQR 87.7-99.9), a high positive predictive value model (positive predictive value 85.7, IQR 57.2-98.2), and the high F_1-score model (F_1=64.4). Our proposed method outperformed 3 classical ML algorithms and an International Classification of Diseases code–based algorithm for fall detection, showing its potential for improved performance in diverse clinical settings.

Conclusions: The developed algorithm provides an automated and accurate method for inpatient fall detection using multidisciplinary progress record notes and a pretrained BERT language model. This method could be implemented in clinical practice to improve patient safety and reduce the occurrence of falls in hospitals.

(JMIR Med Inform 2024;12:e48995) doi:10.2196/48995

KEYWORDS

accidental falls; electronic medical records; data mining; machine learning; patient safety; natural language processing; adverse event

Introduction

Background

Inpatient falls detrimentally impact patients, leading to extended hospital stays and distress among families and caregivers [1-5]. Studies reflect a varying incidence rate of such falls, with 250,000 annually in England and Wales alone [1], and evidence showing 7.5% of patients experience at least 1 fall during hospitalization [2]. Acute care hospitals also report a range of 1 to 9 falls per 1000 bed days, underscoring the pervasive nature of this problem [4]. Patients who fall may experience injuries that increase the risk of comorbidity or even disability [6,7]. They may also experience psychological effects such as anxiety, depression, or loss of confidence, which can affect their recovery and quality of life [8].

Manual chart review is regarded as one of the most common methods to identify inpatient falls [9]. This process involves the thorough examination of patient medical records to gather relevant information on the details of falls. Existing strategies include the Harvard Medical Practice Study [10] and the Global Trigger Tool [11]. Alternative methodologies, such as Patient Safety Indicators, based on International Classification of Diseases (ICD) codes, are used to identify adverse events (AEs), leveraging systematized health care data for detection [12-14]. However, these methodologies, while widely used, present challenges due to the time-consuming nature of ICD coding and manual chart reviews, potentially causing delays in recording and detecting AEs [15,16].

Free text data in electronic medical records (EMRs) offer rich, up-to-date insights into patients' health status, medications, and various narrative content. Despite its wealth of information, the unstructured nature of this data necessitates chart reviews, a labor-intensive process, to identify inpatient falls [17]. There has been an increasing interest in recent years in applying natural language processing (NLP) techniques to electronic clinical notes to automate disease identification and create clinical support decision systems [18-24].

Previous NLP studies in the detection of patient fall including rule-based algorithms [25,26] and machine learning (ML) methods [27-30] have been explored, but they often struggle with the variety and complexity of clinical language.

The deep learning model Bidirectional Encoder Representation from Transformers (BERT) [31] can effectively address these challenges. It uses transformer architecture to understand text contextually, handling linguistic complexity, abbreviations, and data gaps, thereby augmenting text understanding from EMR [20]. The use of transformer-based methods to understand EMR text data has emerged as a promising new trend in automatic clinical text analysis [32].

Objectives

In this study, we intend to pretrain an existing model, BioClinical BERT [33], with free text data from Alberta hospital EMRs to develop an Alberta hospital notes-specific BERT model (AHN-BERT). The pretrained language model would serve as a feature extraction layer in a neural network to identify inpatient falls. We hypothesize that fine-tuning BERT on local hospital data will enable more accurate fall detection compared with generic models. Additionally, we expect AHN-BERT will outperform conventional rule-based and ML approaches, as well as ICD code methods, in detecting falls from unstructured EMR notes in near real time. By evaluating AHN-BERT against current techniques, we hope to demonstrate the value of transfer learning with BERT for improved efficiency and generalizability in surfacing patient safety events from clinical text. Ultimately, our goal is to advance the detection of inpatient falls, allowing for more detailed and accurate patient safety interventions. An improved fall detection system could potentially enable health care providers to swiftly implement preventive measures, reducing the incidence and severity of falls. Additionally, through the facilitation of access and analysis of fall-related data, our system could become an invaluable resource for researchers investigating fall prevention and associated subjects.

Methods

Overview

In our methodology, we emphasized a detailed and transparent approach, covering all aspects from data collection to model validation. This comprehensive process, reflecting best practices in research reporting [34], ensures clarity and precision in our multivariable prediction model, providing an in-depth understanding of its performance and applicability.

Source of Data

Our study is a retrospective analysis. We used a stratified random sample of adult patients admitted to acute care hospitals in Calgary, Alberta. We linked the extracted EMR data to Sunrise Clinical Manager (SCM) records and ICD-coded discharge abstract database (DAD) using an established mechanism [35]. Both tables are stored and managed by the Oracle database.

The chart reviewer team consists of 6 registered nurses with 1 to 10 years of experience using SCM for clinical care. The nurses followed a training procedure, and 1 trained nurse became the project lead for quality assurance. The training involved learning the condition definitions and practicing reviewing each chart systematically. Reviewers examined the entire record for specified conditions and consulted each other with questions. In the process of training and quality assurance, we tested interrater reliability using Conger, Fleiss, and Light κ methods, with 2 nurses reviewing the same set of 10 charts for consensus on AEs. Where agreement was poor ($\kappa<0.60$), retraining occurred until high agreement ($\kappa>0.80$) was achieved [36]. Reviewers then proceeded independently with REDCap tool (Vanderbilt University).

The chart review data served as the reference standard to develop and evaluate our fall detection model. We focused on multidisciplinary progress records (MPRs) for fall detection, as chart review data showed most falls (115/155, 73.7%) were documented in MPRs by nursing staff. We created supervised data sets for the classification task to identify optimal fall detection timing, including 1-day (fall day MRPs notes), 2-day (fall day + day after), 3-day (fall day + 2 days after), and full hospitalization MRPs. All supervised data sets were labeled to

indicate whether notes were associated with inpatient falls. For the training of our model, we used both cases (falls) and controls (nonfalls) at a ratio of 1:29. This was done to ensure the model was exposed to a balanced representation of both scenarios. Our test set mirrored the real-world data distribution to enable an accurate evaluation of model performance. In addition, we constructed an unsupervised corpus specifically for language model pretraining. This corpus comprises free-text note data and does not rely on any predetermined labels or annotations.

Participants

At the time of the study, a total of 4393 charts were reviewed, among which we identified a total of 155 records as falls and 4238 records as no falls. The study included only the first admission of each patient, even if they had multiple hospitalizations within the study period. We exclusively focused on adults 18 to 100 years of age, thereby excluding minors and centenarians. Furthermore, if a patient had multiple fall incidents, only the most recent record was considered, although no such cases were identified during the study. The temporal framework for the study encompassed a decade, from 2010 to 2020. Exclusion criteria were also clearly defined: patients without unstructured note data or those who could not be linked using our established data linkage mechanism were omitted from the study.

Missing Data and Data Cleaning

Our study implemented rigorous data cleaning to ensure data integrity. After conducting a conflict review and excluding records with inconsistencies in fall status documentation (17 records checked for both falls and no falls), failed data linkage (1 record), temporal conflicts between fall and admission dates (5 records), and missing MPR documentations (47 records), the final clean data set totaled 4323 records (142 falls and 4181 no falls).

Outcome and Variables

The desired outcome of our proposed framework is to predict whether a patient's daily progress note contains hints about inpatient falls. The input to our model is each patient's n-day note. We use a BERT model to represent the textual data in numerical format, also known as contextualized word embeddings.

The input text is represented by 768-dimensional feature vectors, which can be considered as 768 variables. However, due to the distributed representation of neural language models, each variable does not represent a single word. Instead, individual variables preserve contextual information segments for each word, constituting meaningful vector representations of the entire input text.

On a related note, we have also collected and analyzed several demographic and clinical variables for our patient cohort from DAD database. Although not directly used in our predictive modeling, these variables furnish invaluable insights into the characteristics of our study population and contribute to the overall richness of our research data. These include age at the time of admission, sex, the incidence of intensive care unit visits during the hospital stays, the length of the hospital stays, and the hospital's geographical location. The latter was particularly focused on 3 acute care hospitals based in Calgary, Alberta: hospitals "A," "B," and "C." These variables help us understand the context in which the patient notes were written and may influence the interpretation of the model's results.

Sample Size

We included all patient data that has been reviewed by the reviewer team and filtered out from the inclusion-exclusion criteria.

Model Development

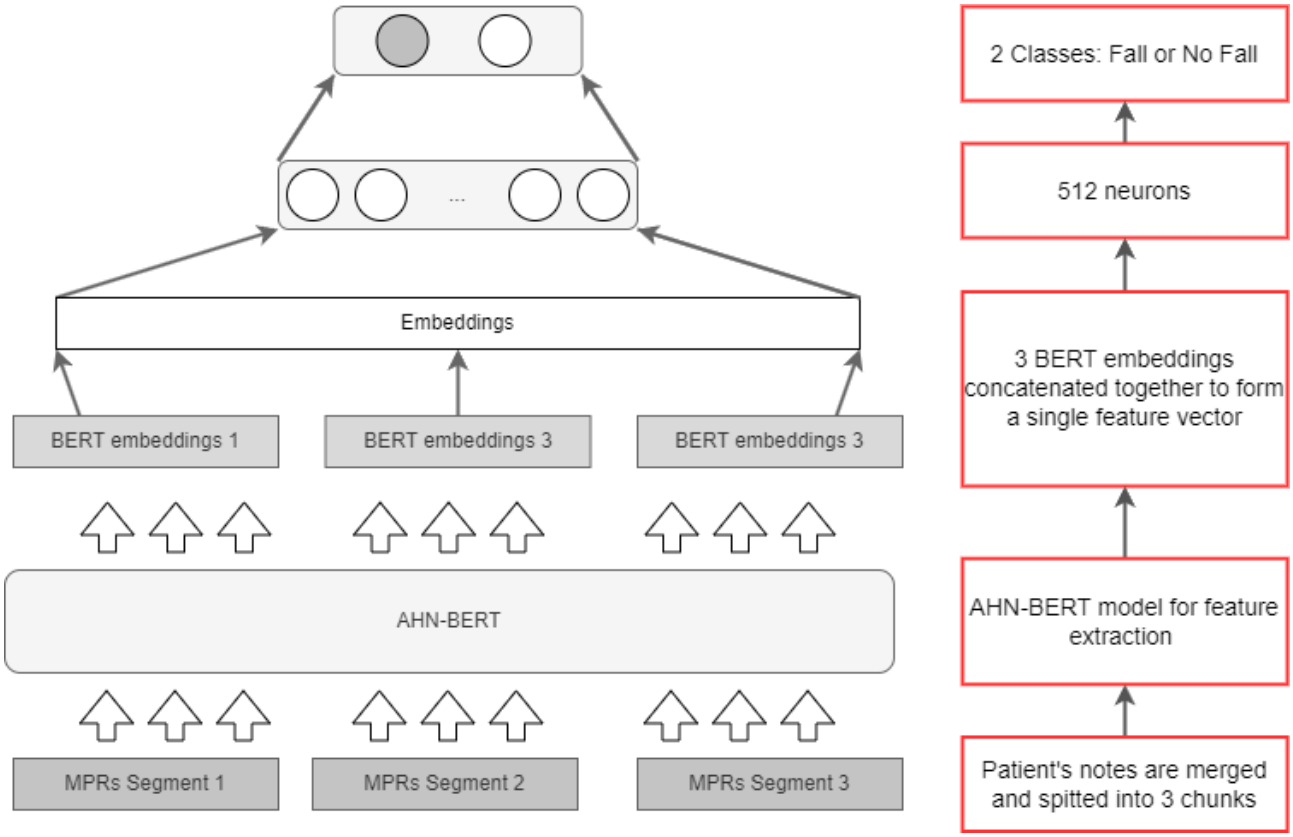
To conform to the BERT input length limit, MPR notes exceeding 400 tokens were programmatically split into segments under 400 words, preserving contextual information. All notes underwent preprocessing including removal of extraneous headers, signatures, whitespaces, and escape characters using regular expressions, and duplicate sentences were eliminated.

Our model architecture comprises 2 key components, an Alberta hospital note-specific BERT model for contextual feature extraction from clinical text, which we term AHN-BERT, and a feedforward neural network classifier to detect falls from the extracted features (as Figure 1).

AHN-BERT was initialized with weights from BioClinical BERT and further pretrained on our corpus of unsupervised hospital notes to adapt to local clinical terminology and language patterns. To prevent bias from overly lengthy documents, notes were filtered to be between 50 and 5000 tokens prior to pretraining. AHN-BERT was pretrained using a masked language modeling technique on 15% of randomly selected input tokens, enabling learning of contextual representations of clinical text without explicit labels. For feature extraction, AHN-BERT processes up to 3 concatenated note segments under 400 tokens. The resulting "[CLS]" vectors summarizing the semantic content of each segment are aggregated via concatenation to represent the full note's contextual information [37].

A feedforward neural network is used as the classifier, taking the concatenated features as input. The network comprises fully connected layers to map the features into class probabilities for fall detection. Dropout regularization is implemented in the classifier to prevent overfitting to the training data. Sigmoid activation in the output layer provides posterior probabilities for the binary fall classification task.

Figure 1. Proposed model architecture. AHN-BERT: Alberta hospital notes-specific BERT; BERT: Bidirectional Encoder Representation from Transformers; MPR: multidisciplinary progress record.



Model Assessment

First, to determine the optimal timeframe for note selection that best represents inpatient falls, we compared model performance using 1-day, 2-day, 3-day, and complete patient note data sets. Since the exact time lapse between an inpatient fall and corresponding documentation is variable, we evaluated these distinct time intervals in a data-driven approach to identify the optimal period for note selection. We used the same model architecture and pretrained AHN-BERT for all data sets, comparing training and validation loss convergence and evaluation metrics to assess performance.

Second, we tuned the classification probability threshold to balance model sensitivity and precision. The threshold denotes the cutoff for determining class membership based on predicted probabilities. By optimizing the threshold, we controlled the tradeoff between correctly identifying true positives and avoiding false positives. We developed three distinct models by threshold tuning for different purposes: (1) a high-sensitivity model that maximizes sensitivity by optimizing the threshold, (2) a high positive predictive value (PPV) model that maximizes PPV through threshold optimization, and (3) a high F_1-score model that balances sensitivity and PPV by optimizing the threshold, serving as a general-purpose model.

Third, we conducted a comparative evaluation between our top-performing neural network model and several other approaches, including 2 alternative BERT-based models, 3 conventional ML models, and an ICD-code–based algorithm. The 2 additional BERT-based models used original pretrained BERT and BioClinical BERT as feature extractors. For the 3 conventional ML models (support vector machine, logistic regression, and decision tree classifiers), we used bag-of-words features and term frequency-inverse document frequency weighting. These models were trained and compared on the 1-day MPRs data set. The ICD-code algorithm was applied to the same patient cohort but relied on administrative diagnosis codes rather than clinical notes. It aimed to demonstrate the efficacy of standard diagnostic codes for identifying falls compared with our neural network model. Falls were identified by the presence of ICD-10 codes W00-W20 when not listed as the primary diagnosis.

Statistical Analysis

The characteristics of the patients included in the study were thoroughly evaluated. These characteristics encompassed age, sex, the incidence of intensive care unit visits, the length of their hospital stay, and their originating hospitals. We summarized categorical variables as frequencies and percentages, while continuous variables were expressed as medians and IQRs. The χ^2 test was used for categorical variables to determine statistical differences, while the Wilcoxon rank-sum test was used for continuous variables. A P value threshold of 5% or lower was set to denote statistical significance.

To evaluate our ML model, we calculated several statistical metrics such as sensitivity, specificity, PPV, negative predictive value, accuracy, and F_1-score.

Computational Environment

Our study harnessed a high-performance computing environment, primarily driven by an NVIDIA GeForce RTX

3080 GPU with 16GB of memory, vital for pretraining and fine-tuning our language model. The statistical analysis and experiment leveraged Python 3.8, and libraries such as NumPy [38], Scikit-learn [39], Pandas [40], and PyTorch [41] for tasks like data processing and modeling.

Ethics Approval

This study was approved by the Conjoint Health Research Ethics Board at the University of Calgary (REB21-0416). Patient consent was waived as part of the ethics board review process.

Results

Participants

Our final study cohort contains 4323 individuals, with 142 (3.28%) patients identified by chart reviewers as having falls recorded in their medical charts during their hospital stay. The remaining 4181 (96.7%) did not fall. All patients were successfully linked to the SCM and DAD by unique identification number and admission date. Table 1 presents the descriptive statistics in general. Multimedia Appendix 1 further stratifies Table 1 into respective hospitals (Multimedia Appendix 1).

Table 1. Descriptive statistics of the study cohort.

	Total (n=4323)	Confirmed fall (n=142)[a]	No fall (n=4181)[b]	*P* value[c]
Age, median (IQRs[d])	62.0 (48.5-75.5)	71.0 (59.6-82.4)	61.0 (47.5-74.5)	<001
Sex (male), n (%)	2169 (50.2)	73 (51.4)	2096 (50.1)	.77
ICU[e] visit, n (%)	163 (3.8)	19 (13.4)	144 (3.4)	<.001
Length of hospital stay (days), median (IQRs)	3.0 (0.5-5.5)	12.0 (1.5-22.5)	3.0 (0.5-5.5)	<.001
Hospitals, n (%)[f]				.04
Hospital “A”	3548 (82.1)	104 (73.2)	3444 (82.4)	
Hospital “B”	651 (15.1)	34 (23.9)	617 (14.8)	
Hospital “C”	124 (2.8)	4 (2.9)	120 (2.8)	

[a]A term used in the study to refer to patients who fell during their hospital stay and were confirmed to have fallen through medical records or other documentation.

[b]A term used in the study to refer to patients who did not fall during their hospital stay.

[c]A measure indicating the statistical significance (P<.05) of the observed difference between groups.

[d]A measure of statistical dispersion representing the difference between the 75th and 25th percentiles of a data set.

[e]ICU: intensive care unit.

[f]Three different hospitals were included in the study: hospitals A, B, and C.

Model and Framework Assessment

First, to determine the optimal timeframe, we compared 1-day, 2-day, 3-day, and complete note data sets using the same model architecture and AHN-BERT pretrained embeddings. We trained each model for 200 epochs, with the primary goal of comparing their overall performance on the test sets. Evaluating performance metrics and training convergence, the 1-day data set was most effective and efficient, achieving 93.0% sensitivity and 83.0% specificity.

Second, we optimized the classification threshold to balance sensitivity and precision. These models maximize sensitivity, PPV, and F_1-score respectively. As results are shown in Table 2, our proposed architecture with AHN-BERT achieved overall the highest metrics among the comparison.

Third, the comparative assessment showed our approach outperformed 2 alternative BERT models, 3 classical ML models (support vector machine, logistic regression, and decision tree), and an ICD-code algorithm. The BERT models used original BERT and BioClinical BERT embeddings, while the ML models used bag-of-words and term frequency-inverse document frequency on the 1-day data set. The ICD method relied on administrative codes rather than text. Our neural network model demonstrated superior inpatient fall detection across different methods and data sources.

Our high sensitivity model exhibited 97.7% sensitivity, enabling near-perfect capture of relevant notes, along with 82.3% accuracy, but a low 26.8% F_1-score. The high PPV model achieved 97.5% accuracy, 85.7 % PPV, and 27.9 % sensitivity. The high F_1-model balanced 66.7% sensitivity and 60.5% PPV to optimize 64.4% F_1-score and 97.7% accuracy. In comparison, the ICD-based method had 27.9% sensitivity, while traditional classifiers achieved 51.2%-76.7% sensitivities and 8.3-15.8 PPVs.

The result of the probability-based threshold adjustment in accordance with PPV, sensitivity, and F_1-score is shown in Figure 2. By adjusting the classification threshold, we can control the trade-off between sensitivity and precision (as Figure 3).

Table 2. Performance of proposed deep learning models, classical machine learning methods, and International Classification of Diseases–based algorithms on fall identification with 1-day data set.

Category and model name	Sensitivity (%), (95% CI)	Specificity (%), (95% CI)	PPV[a] (%), (95% CI)	NPV[b] (%), (95% CI)	Accuracy (%), (95% CI)	F_1-score[c] (%)
BERT[d]-based models						
AHN-BERT[e] (high sensitivity)	97.7 (87.7-99.9)	81.8 (79.6-83.9)	15.6 (14.0-17.3)	99.9 (99.3-100.0)	82.3 (80.1-84.4)	26.8
AHN-BERT (high PPV)	27.9 (15.3-43.7)	99.8 (99.4-100.0)	85.7 (57.2-98.2)	97.6 (96.6-98.4)	97.5(96.5-98.2)	42.1
AHN-BERT (high F_1)	66.7 (49.8-80.9)	99.0 (98.2-99.5)	60.5 (44.4-75.0)	98.7 (97.8-99.2)	97.7 (96.7-98.4)	63.4
BERT-uncased	79.1 (64.0-9 0.0)	61.4 (58.6-64.1)	6.6 (5.6-7.7)	98.8 (97.9-99.4)	62.0 (59.3-64.6)	12.1
BioClinical BERT	74.4 (58.8-86.5)	69.8 (67.2-72.4)	7.8 (6.5-9.3)	98.8 (97.9-99.3)	70.0 (67.4-72.5)	14.1
Classical machine learning classifier						
Support vector machine	76.7 (61.4-88.2)	85.0 (82.9-86.9)	14.9 (12.5-17.8)	99.1 (98.4-99.5)	84.7 (82.6-86.6)	25.0
Logistic regression	74.4 (58.8-86.5)	86.4 (84.3-88.2)	15.8 (13.0-19.0)	99.0 (98.3-99.4)	86.0 (83.9-87.8)	26.0
Decision tree	51.2 (35.5-66.7)	93.9 (92.4-95.1)	22.2 (14.5-31.7)	98.3 (97.3-98.9)	92.4 (90.9-93.8)	31.0
Rule-based classifier						
ICD 10[f]	27.9 (15.3-43.7)	92.3 (90.7-93.8)	11.1 (6.9-17.3)	97.4 (96.9-97.8)	90.2 (88.5-91.8)	15.9

[a]PPV: positive predictive value. The proportion of true positive results among all positive results.

[b]NPV: negative predictive value. The proportion of true negative results among all negative results.

[c]F_1-score: a measure of a model's accuracy that considers both sensitivity and PPV.

[d]BERT: Bidirectional Encoder Representations from Transformers.

[e]AHN-BERT: Alberta hospital notes-specific BERT.

[f]ICD 10: International Classification of Diseases, 10th Revision.

Figure 2. Performance metrics at varying thresholds: PPV, sensitivity, and F1-score. PPV: positive predictive value.

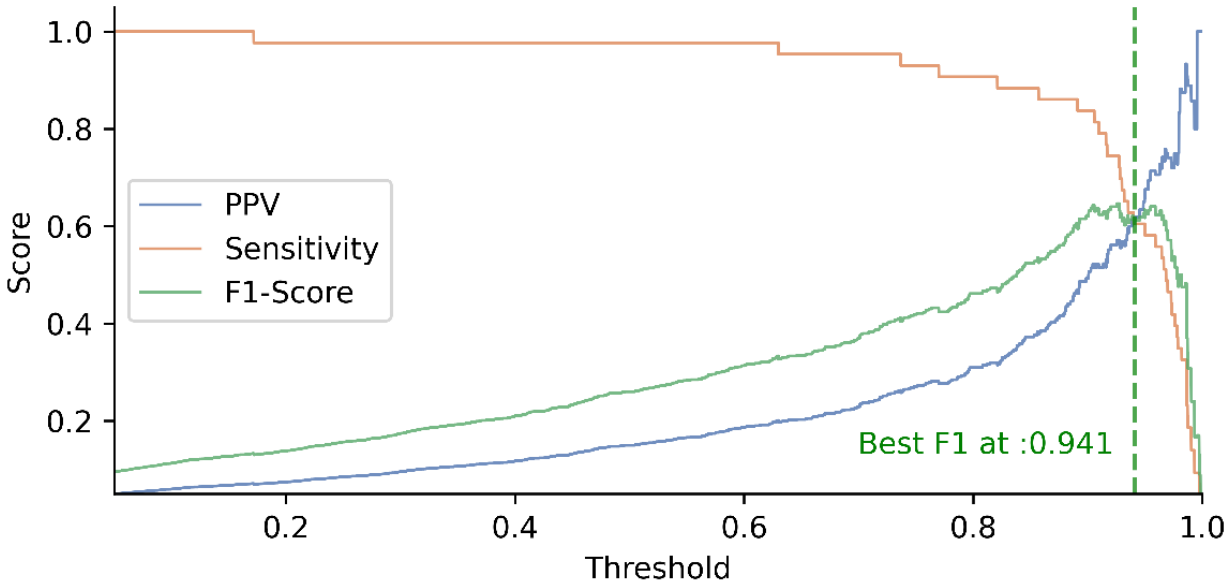
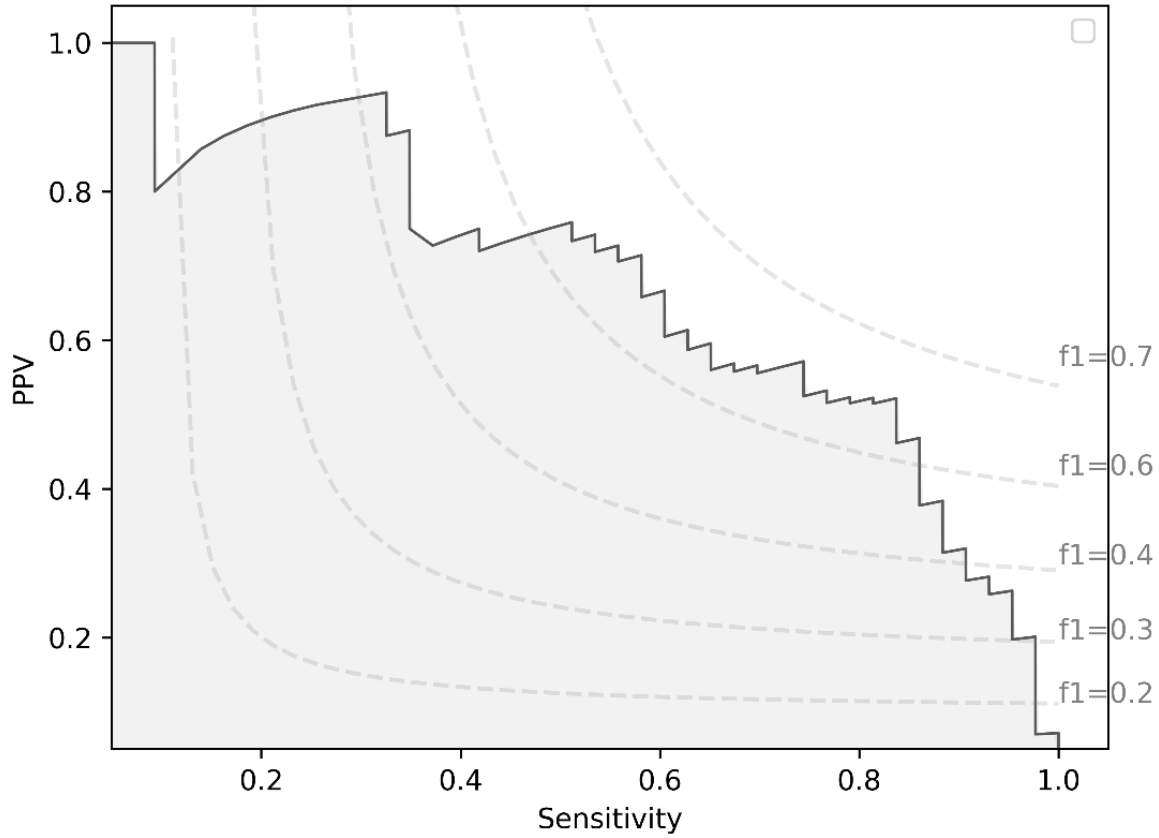


Figure 3. Percision-recall curve. PPV: positive predictive value.



Discussion

Principal Results

In our study, we illustrate how a BERT-based model substantially outperforms the ICD-based algorithm in fall detection within the hospital setting. This superiority stems from the model's ability to process EMR text data in real time, enabling rapid identification of falls. In contrast, ICD codes are assigned retrospectively, leading to delays in fall detection and intervention. Using BERT's advanced NLP facilitates accurate, efficient, and generalizable analysis of clinical notes for surveillance applications.

Specifically, our proposed AHN-BERT model surpasses generic BERT, conventional ML, and ICD-codes. Fine-tuning BERT on local hospital notes better captures local domain-specific language and context, boosting performance. Additionally, combining high-sensitivity and high-PPV models enables optimized 2-stage fall detection by adjusting decision thresholds to balance false positives and negatives. This provides flexibility for different use cases and challenging tasks.

Furthermore, our study provides valuable insights into the optimal time frame for defining falls empirically. This comparison sheds light on the potential benefits of using a finer-grained time interval, which could improve the generalizability and applicability of the model across different populations and settings. Understanding the optimal period for detecting fall incidents can guide the development and implementation of targeted public health interventions. Health surveillance data can be used to evaluate the effectiveness of these interventions and inform future strategies for fall prevention and management. By determining the most suitable time interval for defining fall incidents, health surveillance systems can better allocate resources to areas with a higher risk of falls. This may result in more efficient and effective public health efforts, improving health outcomes for at-risk populations.

Applications

Our models leverage unstructured EMR data to accurately detect inpatient falls, enabling health care systems to enact tailored prevention measures and reduce fall-associated injuries. The automation of extensive clinical documentation review accelerates health care surveillance and quality improvement processes.

Regarding research applications, our algorithms can extract comprehensive fall data from EMR text to support developing evidence-based interventions.

The proposed framework has broad applicability beyond fall detection for tasks like diagnosis prediction, medication adherence monitoring, and adverse drug event identification. This adaptability improves health care outcomes, patient safety, and quality of care.

Strength and Limitations

In our research, the AHN-BERT model has shown remarkable superiority over traditional ICD-based algorithms in fall detection within hospital environments. This enhanced performance is primarily attributed to the model's proficiency in processing and understanding the nuances of EMRs text data. Unlike ICD codes, which can sometimes result in undercoding

or loss of information, the nursing notes processed by our model are more closely aligned with the actual circumstances of inpatient falls. The ability of AHN-BERT to immediately and accurately process this data is a substantial advancement, ensuring that fall detection is not only more precise but also more reflective of the true clinical scenario. Additionally, the combination of high-sensitivity and high-PPV models in our 2-stage fall detection system allows for adjustable decision thresholds, thus balancing false positives and negatives and providing flexibility across different scenarios.

However, the model faces challenges in balancing high sensitivity with a high PPV due to the imbalanced nature of clinical data. The rarity of AEs like falls leads to a higher rate of false positives, as seen in our data set with a significant imbalance ratio. Our test data set, characterized by a significant 29:1 imbalance, aligns more closely with real-world clinical scenarios than balanced data sets used in some prior studies [27], which, while yielding promising results, may not fully represent practical conditions. This intentional choice ensures that our model's performance is tested under conditions typical of rare events like falls, thereby enhancing its relevance and utility in actual clinical settings.

Second, the effectiveness of our models depends on the quality and comprehensiveness of documentation. If fall events or associated risk factors are not well documented, our model, like any data-driven model, may have difficulty detecting them. This underscores the importance of careful, detailed clinical documentation to enhance the effectiveness of monitoring applications. In addition, our study also assumes a certain level of linguistic and terminological consistency within the EMR data. Variations in documentation styles across different health care providers could potentially impact the model's performance, suggesting that future models should incorporate strategies, for example, pretraining the ML, to mitigate such discrepancies. Last, the differentiation between a history of falls and inpatient falls presents a challenge, as it could potentially lead to false positive predictions if falls that occurred prior to hospitalization are documented in the notes. Although the BERT model's contextual understanding can partially alleviate this issue, we acknowledge that more improvements are needed. As part of our future work, we aim to further refine our model to better handle such complexities.

Conclusions

This study developed and evaluated BERT-based NLP models for the automated detection of falls from electronic clinical notes. The developed models provided a more accurate and timely way to detect falls than traditional ML and ICD-codes–based methods. Moreover, we provided a masked language model technique to pretrain a pre-existing BERT model using clinical text data gathered from various health care facilities in Calgary, Alberta, creating a more local institution-specific and effective AHN-BERT model. By using self-supervised language modeling strategies, we can bypass steps that were regarded as vital in standard ML methods, such as the necessity for thorough text preprocessing, complex feature engineering, and a considerable amount of labeled data. In addition, by exploring the optimal period for fall incident detection and selecting 1-day notes for our final architecture, our model contributes to enhanced patient safety and care with less noise.

Acknowledgments

YX and CAE received research support funding from Canadian Institutes of Health Research through grant number DC0190GP. GW was supported by the Canadian Institutes of Health Research postdoctoral fellowship, O'Brien Institute for Public Health Postdoctoral Scholarship, Cumming School of Medicine Postdoctoral Scholarship at the University of Calgary, and the Network of Alberta Health Economists Postdoctoral Fellowship at the University of Alberta.

Authors' Contributions

YX, CAE, HQ, GW, and CC were responsible for the study planning, conceptualization, and coordination. SL, EAM, and GW managed data retrieval, data linkage, and data quality assurance. The design and development of the neural network architecture were carried out by CC and YX. CC conducted clinical note preprocessing, analysis, and model evaluation. The reference standard development and chart review study design was executed by YX, CAE, NS, DAS, and GW. SL and CC drafted the manuscript and GW drafted the Methods (Study Cohort and Data Sources). GW, JP, DAS, EAM, CAE, NS, and YX participated in discussions and provided comments on the manuscript. All authors contributed to the revision and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Descriptive statistics for the study cohort from each hospital.
[DOCX File , 19 KB - medinform_v12i1e48995_app1.docx]

References

1. Morris R, O'Riordan S. Prevention of falls in hospital. Clin Med (Lond) 2017;17(4):360-362 [FREE Full text] [doi: 10.7861/clinmedicine.17-4-360] [Medline: 28765417]

2. Schwendimann R, Bühler H, De Geest S, Milisen K. Characteristics of hospital inpatient falls across clinical departments. Gerontology 2008;54(6):342-348. [doi: 10.1159/000129954] [Medline: 18460867]
3. Zeneli A, Montalti S, Masciangelo I, Manieri G, Golinucci M, Nanni O, et al. Fall predictors in hospitalized patients living with cancer: a case-control study. Support Care Cancer 2022;30(10):7835-7843. [doi: 10.1007/s00520-022-07208-x] [Medline: 35705752]
4. Oliver D, Healey F, Haines TP. Preventing falls and fall-related injuries in hospitals. Clin Geriatr Med 2010;26(4):645-692. [doi: 10.1016/j.cger.2010.06.005] [Medline: 20934615]
5. Morello RT, Barker AL, Watts JJ, Haines T, Zavarsek SS, Hill KD, et al. The extra resource burden of in-hospital falls: a cost of falls study. Med J Aust 2015;203(9):367. [doi: 10.5694/mja15.00296] [Medline: 26510807]
6. Miake-Lye IM, Hempel S, Ganz DA, Shekelle PG. Inpatient fall prevention programs as a patient safety strategy: a systematic review. Ann Intern Med 2013;158(5 Pt 2):390-396 [FREE Full text] [doi: 10.7326/0003-4819-158-5-201303051-00005] [Medline: 23460095]
7. King B, Pecanac K, Krupp A, Liebzeit D, Mahoney J. Impact of fall prevention on nurses and care of fall risk patients. Gerontologist 2018;58(2):331-340 [FREE Full text] [doi: 10.1093/geront/gnw156] [Medline: 28011591]
8. Rubenstein LZ. Falls in older people: epidemiology, risk factors and strategies for prevention. Age Ageing 2006;35(Suppl 2):ii37-ii41 [FREE Full text] [doi: 10.1093/ageing/afl084] [Medline: 16926202]
9. Murff HJ, Patel VL, Hripcsak G, Bates DW. Detecting adverse events for patient safety research: a review of current methodologies. J Biomed Inform 2003;36(1-2):131-143 [FREE Full text] [doi: 10.1016/j.jbi.2003.08.003] [Medline: 14552854]
10. Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. N Engl J Med 1991 Feb 07;324(6):370-376. [doi: 10.1056/NEJM199102073240604] [Medline: 1987460]
11. Griffin FA, Resar RK. IHI Innovation Series white paper. IHI Global Trigger Tool for Measuring Adverse Events (Second Edition). Cambridge, MA: Institute for Healthcare Improvement; 2009. URL: https://www.ihi.org/ [accessed 2023-03-15]
12. Southern DA, Burnand B, Droesler SE, Flemons W, Forster AJ, Gurevich Y, et al. Deriving ICD-10 codes for patient safety indicators for large-scale surveillance using administrative hospital data. Med Care 2017;55(3):252-260. [doi: 10.1097/MLR.0000000000000649] [Medline: 27635599]
13. Menendez ME, Ring D, Jawa A. Inpatient falls after shoulder arthroplasty. J Shoulder Elbow Surg 2017;26(1):14-19. [doi: 10.1016/j.jse.2016.06.008] [Medline: 27522341]
14. Memtsoudis SG, Danninger T, Rasul R, Poeran J, Gerner P, Stundner O, et al. Inpatient falls after total knee arthroplasty: the role of anesthesia type and peripheral nerve blocks. Anesthesiology 2014;120(3):551-563 [FREE Full text] [doi: 10.1097/ALN.0000000000000120] [Medline: 24534855]
15. Schroll JB, Maund E, Gøtzsche PC. Challenges in coding adverse events in clinical trials: a systematic review. PLoS One 2012;7(7):e41174 [FREE Full text] [doi: 10.1371/journal.pone.0041174] [Medline: 22911755]
16. Golder S, Loke YK, Wright K, Norman G. Reporting of adverse events in published and unpublished studies of health care interventions: a systematic review. PLoS Med 2016;13(9):e1002127 [FREE Full text] [doi: 10.1371/journal.pmed.1002127] [Medline: 27649528]
17. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng 2018;2018:4302425 [FREE Full text] [doi: 10.1155/2018/4302425] [Medline: 29849998]
18. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885 [FREE Full text] [doi: 10.1136/bmj.h1885] [Medline: 25911572]
19. 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;7(1):e13039 [FREE Full text] [doi: 10.2196/13039] [Medline: 30862607]
20. 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;8(11):e22508 [FREE Full text] [doi: 10.2196/22508] [Medline: 33245284]
21. Arnaud É, Elbattah M, Gignon M, Dequen G. Learning embeddings from free-text triage notes using pretrained transformer models. 2022 Presented at: Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies—Volume 5 HEALTHINF: Scale-IT-up; February 9-11, 2022; Vienna, Austria p. 835-841. [doi: 10.5220/0011012800003123]
22. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: a review. Trends Anaesth Crit Care 2021;38:4-9. [doi: 10.1016/j.tacc.2021.02.007]
23. Yan H, Rahgozar A, Sethuram C, Karunananthan S, Archibald D, Bradley L, et al. Natural language processing to identify digital learning tools in postgraduate family medicine: protocol for a scoping review. JMIR Res Protoc 2022;11(5):e34575 [FREE Full text] [doi: 10.2196/34575] [Medline: 35499861]

24. Gaviria-Valencia S, Murphy SP, Kaggal VC, McBane Ii RD, Rooke TW, Chaudhry R, et al. 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 [FREE Full text] [doi: 10.2196/40964] [Medline: 36826984]
25. Dolci E, Schärer B, Grossmann N, Musy SN, Zúñiga F, Bachnick S, et al. Automated fall detection algorithm with global trigger tool, incident reports, manual chart review, and patient-reported falls: algorithm development and validation with a retrospective diagnostic accuracy study. J Med Internet Res 2020;22(9):e19516 [FREE Full text] [doi: 10.2196/19516] [Medline: 32955445]
26. Toyabe SI. Detecting inpatient falls by using natural language processing of electronic medical records. BMC Health Serv Res 2012;12:448 [FREE Full text] [doi: 10.1186/1472-6963-12-448] [Medline: 23217016]
27. Fu S, Thorsteinsdottir B, Zhang X, Lopes GS, Pagali SR, LeBrasseur NK, et al. A hybrid model to identify fall occurrence from electronic health records. Int J Med Inform 2022;162:104736 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104736] [Medline: 35316697]
28. Jung H, Park HA, Hwang H. Improving prediction of fall risk using electronic health record data with various types and sources at multiple times. Comput Inform Nurs 2020;38(3):157-164. [doi: 10.1097/CIN.0000000000000561] [Medline: 31498252]
29. Nakatani H, Nakao M, Uchiyama H, Toyoshiba H, Ochiai C. Predicting inpatient falls using natural language processing of nursing records obtained from Japanese electronic medical records: case-control study. JMIR Med Inform 2020;8(4):e16970 [FREE Full text] [doi: 10.2196/16970] [Medline: 32319959]
30. Thapa R, Garikipati A, Shokouhi S, Hurtado M, Barnes G, Hoffman J, et al. Predicting falls in long-term care facilities: machine learning study. JMIR Aging 2022;5(2):e35373 [FREE Full text] [doi: 10.2196/35373] [Medline: 35363146]
31. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), Vol 1; June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186.
32. 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;9(1):e23086 [FREE Full text] [doi: 10.2196/23086] [Medline: 33480858]
33. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop.: Association for Computational Linguistics; 2019 Presented at: The 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, Minnesota, USA p. 72-78. [doi: 10.18653/v1/W19-1909]
34. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015;162(10):735-736 [FREE Full text] [doi: 10.7326/L15-5093-2] [Medline: 25984857]
35. Lee S, Xu Y, Apos Souza AGD, Martin EA, Doktorchik C, Zhang Z, et al. Unlocking the potential of electronic health records for health research. Int J Popul Data Sci 2020;5(1):1123 [FREE Full text] [doi: 10.23889/ijpds.v5i1.1123] [Medline: 32935049]
36. Eastwood CA, Southern DA, Khair S, Doktorchik C, Cullen D, Ghali WA, et al. Field testing a new ICD coding system: methods and early experiences with ICD-11 beta version 2018. BMC Res Notes 2022;15(1):343 [FREE Full text] [doi: 10.1186/s13104-022-06238-2] [Medline: 36348430]
37. Qiao Y, Xiong C, Liu Z, Liu Z. Understanding the behaviors of BERT in ranking. arXiv :1-4 Preprint posted online on April 16, 2019. [FREE Full text] [doi: 10.1090/mbk/121/79]
38. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature 2020;585(7825):357-362 [FREE Full text] [doi: 10.1038/s41586-020-2649-2] [Medline: 32939066]
39. 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 [FREE Full text]
40. McKinney W. Data structures for statistical computing in Python. 2010 Presented at: Proceedings of the 9th Python in Science Conference (SciPy 2010); June 28-30, 2010; Austin, TX. [doi: 10.25080/majora-92bf1922-00a]
41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 2019;32:1-12 [FREE Full text]

Abbreviations

AE: adverse event
AHN-BERT: Alberta hospital notes-specific BERT
BERT: Bidirectional Encoder Representation from Transformers
DAD: discharge abstract database
EMR: electronic medical record
ICD: International Classification of Diseases
ML: machine learning

MPR: multidisciplinary progress record
NLP: natural language processing
PPV: positive predictive value
SCM: Sunrise Clinical Manager

Edited by C Lovis; submitted 15.05.23; peer-reviewed by S Musy, M Elbattah; comments to author 05.07.23; revised version received 24.07.23; accepted 23.12.23; published 30.01.24.

Please cite as:
Cheligeer C, Wu G, Lee S, Pan J, Southern DA, Martin EA, Sapiro N, Eastwood CA, Quan H, Xu Y
BERT-Based Neural Network for Inpatient Fall Detection From Electronic Medical Records: Retrospective Cohort Study
JMIR Med Inform 2024;12:e48995
URL: https://medinform.jmir.org/2024/1/e48995
doi:10.2196/48995
PMID:38289643

Original Paper

Mining Clinical Notes for Physical Rehabilitation Exercise Information: Natural Language Processing Algorithm Development and Validation Study

Sonish Sivarajkumar[1], BS; Fengyi Gao[2], MS; Parker Denny[3], DPT; Bayan Aldhahwani[3,4], MS, PT; Shyam Visweswaran[1,5,6], MD, PhD; Allyn Bove[3], DPT, PhD; Yanshan Wang[1,2,5,6], PhD

[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, United States

[2]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, United States

[3]Department of Physical Therapy, University of Pittsburgh, Pittsburgh, PA, United States

[4]Department of Physical Therapy, Umm Al-Qura University, Makkah, Saudi Arabia

[5]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

[6]Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
Yanshan Wang, PhD
Department of Health Information Management
University of Pittsburgh
6026 Forbes Tower
Pittsburgh, PA, 15260
United States
Phone: 1 4123832712
Email: yanshan.wang@pitt.edu

Abstract

Background: The rehabilitation of a patient who had a stroke requires precise, personalized treatment plans. Natural language processing (NLP) offers the potential to extract valuable exercise information from clinical notes, aiding in the development of more effective rehabilitation strategies.

Objective: This study aims to develop and evaluate a variety of NLP algorithms to extract and categorize physical rehabilitation exercise information from the clinical notes of patients who had a stroke treated at the University of Pittsburgh Medical Center.

Methods: A cohort of 13,605 patients diagnosed with stroke was identified, and their clinical notes containing rehabilitation therapy notes were retrieved. A comprehensive clinical ontology was created to represent various aspects of physical rehabilitation exercises. State-of-the-art NLP algorithms were then developed and compared, including rule-based, machine learning–based algorithms (support vector machine, logistic regression, gradient boosting, and AdaBoost) and large language model (LLM)–based algorithms (ChatGPT [OpenAI]). The study focused on key performance metrics, particularly F_1-scores, to evaluate algorithm effectiveness.

Results: The analysis was conducted on a data set comprising 23,724 notes with detailed demographic and clinical characteristics. The rule-based NLP algorithm demonstrated superior performance in most areas, particularly in detecting the "Right Side" location with an F_1-score of 0.975, outperforming gradient boosting by 0.063. Gradient boosting excelled in "Lower Extremity" location detection (F_1-score: 0.978), surpassing rule-based NLP by 0.023. It also showed notable performance in the "Passive Range of Motion" detection with an F_1-score of 0.970, a 0.032 improvement over rule-based NLP. The rule-based algorithm efficiently handled "Duration," "Sets," and "Reps" with F_1-scores up to 0.65. LLM-based NLP, particularly ChatGPT with few-shot prompts, achieved high recall but generally lower precision and F_1-scores. However, it notably excelled in "Backward Plane" motion detection, achieving an F_1-score of 0.846, surpassing the rule-based algorithm's 0.720.

Conclusions: The study successfully developed and evaluated multiple NLP algorithms, revealing the strengths and weaknesses of each in extracting physical rehabilitation exercise information from clinical notes. The detailed ontology and the robust performance of the rule-based and gradient boosting algorithms demonstrate significant potential for enhancing precision

rehabilitation. These findings contribute to the ongoing efforts to integrate advanced NLP techniques into health care, moving toward predictive models that can recommend personalized rehabilitation treatments for optimal patient outcomes.

(JMIR Med Inform 2024;12:e52289) doi:10.2196/52289

KEYWORDS

natural language processing; electronic health records; rehabilitation; physical exercise; ChatGPT; artificial intelligence; stroke; physical rehabilitation; rehabilitation therapy; exercise; machine learning

Introduction

Precision medicine is a promising field of research that aims to provide personalized treatment plans for patients [1]. Recent years have seen a rise in interest in this field, as advances in machine learning and data collection techniques have greatly facilitated this research [2]. However, the principles of precision medicine have primarily been applied to the development of medications, and relatively little research has been conducted on their applications in other areas [3]. For instance, although rehabilitation clinics require individualized treatment procedures for patients, little research has been conducted on methods that use data analysis and machine learning to facilitate the design of such procedures [4]. Although the application of precision medicine to physical therapy has proven effective in improving the health of patients, current methods of creating personalized treatments rarely use automated approaches to facilitate decision support [5]. Thus, there is a need for tools to assist in the development of personalized treatments in physical therapy [6]. In the treatment of patients who had a stroke, the lack of decision support tools is especially pronounced, as the available treatments for this condition have not led to consistent outcomes across patient populations [7].

To develop decision support tools for the design of precision rehabilitation treatments for patients who had a stroke, it would be necessary to use electronic health record data to develop a predictive model of existing treatment options and their impact on patient outcomes [8]. However, physical therapy procedures are typically described in unstructured clinical notes, meaning that simple data extraction methods such as database queries cannot be applied to obtain sufficient information. Additionally, the language used to describe these procedures can differ between clinicians, locations, and periods [9]. More advanced natural language processing (NLP) algorithms are required to extract this information from clinical notes, but such a method has not yet been developed for this application.

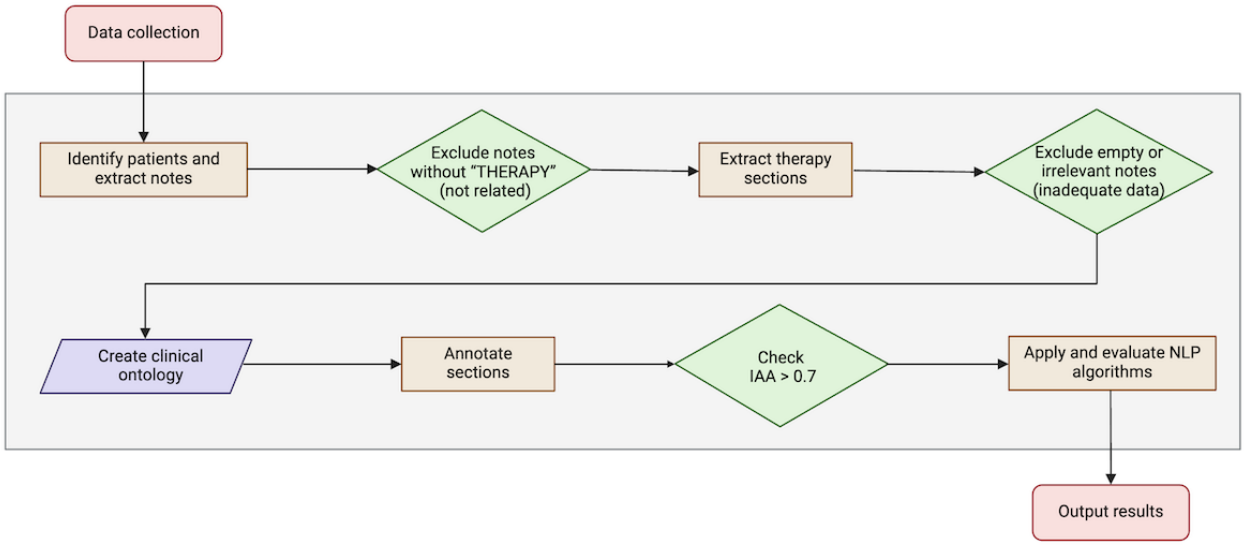
In this paper, we aim to develop and evaluate NLP algorithms to extract physical rehabilitation exercise information from the clinical notes in the electronic health record. Our specific contributions are as follows. First, we created a novel and comprehensive clinical ontology to represent physical rehabilitation exercise information, which includes the type of motion, side of the body, location on the body, the plane of motion, duration, information on sets and reps, exercise purpose, exercise type, and body position. Second, we developed and compared a variety of NLP algorithms leveraging state-of-the-art techniques, including rule-based NLP algorithms, machine learning–based NLP algorithms (ie, support vector machine [SVM], logistic regression [LR], gradient boosting, and AdaBoost), and large language model (LLM)–based NLP algorithms (ie, ChatGPT [OpenAI] [10]) for the extraction of physical rehabilitation exercise from clinical notes. We are among the first to evaluate the capabilities of ChatGPT in extracting useful information from clinical notes.

Methods

Overview

Figure 1 illustrates the data flow and the various stages of the research process. Each of these stages will be described in detail in the following sections.

Figure 1. Flowchart illustrating the data flow throughout the study. IIA: interannotator agreement (IAA); NLP: natural language processing.



Data Collection

The study identified a cohort of patients diagnosed with stroke between January 1, 2016, and December 31, 2016, at University of Pittsburgh Medical Center (UPMC). For these patients, clinical encounter notes created between January 1, 2016, and December 31, 2018, were extracted from the institutional data warehouse. Table 1 provides the demographic characteristics of the patients included in this data set.

Table 1. Demographic information of patients included in the unfiltered data set (N=13,605).

Demographics	Values
Age (years), mean (SD)	75 (16)
Gender, n (%)	
Female	6931 (51)
Male	6673 (49)
Race, n (%)	
Asian	64 (0.5)
Black	1325 (9.7)
White	11,661 (86)
Other	153 (1.1)
Not specified	402 (3)
Ethnicity, n (%)	
Hispanic or Latinx	64 (0.5)
Not Hispanic or Latinx	12,471 (92)
Not specified	984 (7.2)

Ethical Considerations

The study was approved by the University of Pittsburgh's institutional review board (#21040204).

Clinical Ontology for Physical Rehabilitation Exercise

To determine the relevance and hierarchy of extracted information, we developed a clinical ontology consisting of 9 categories of concepts relating to exercise descriptions, informed by consultation with clinical experts (PD, BA, and AB) in the field of physical therapy. In developing our clinical ontology, we also consulted established frameworks such as the International Classification of Functioning, Disability, and Health (ICF) [11] and the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [12]. These comprehensive systems offered valuable insights into the structuring and categorization of health-related concepts, which we adapted for the specific context of physical rehabilitation exercises. Additionally, our ontology incorporates principles from the Unified Medical Language System (UMLS) [13] to ensure compatibility and interoperability with other health care informatics systems.

Each category was given a set of values, as well as examples of how those values might be expressed in clinical notes. The categories are type of motion, side of the body, location on the body, the plane of motion, duration, information on sets and reps, exercise purpose, exercise type, and body position. The ontology also includes examples of indications that the mentioned exercise was not performed during the visit corresponding to the clinical note. This ontology was used to inform both the structure of the annotations and the methods used to extract relevant documents from the data set.

The ontology reflects the complexity and nuance of physical rehabilitation exercises by incorporating terms and categories that are sensitive to the variations and specificities observed in clinical settings. This approach ensures that the ontology not only represents the theoretical model of rehabilitation exercises but also aligns with the practical, real-world application and documentation by health care professionals. Table 2 displays the 9 categories for 3 exercise descriptions (performed in-office, home exercise program, and not performed), with sets and reps split into separate rows and including negations and out-of-office exercises at the bottom.

Table 2. Summary of the clinical ontology used for annotations.

Category	Data type	Concepts
Exercise description	Enumerated	Performed in-office, home exercise program, not performed
Type of motion	Enumerated	ROM[a], active ROM, active-assisted ROM, and passive ROM
Side of body	Enumerated	Right, left, bilateral, unilateral, contralateral, and ipsilateral
Location on body	Enumerated	Upper extremity (arms), lower extremity (legs), hip, thigh, knee, ankle, foot, heel, toe, shoulder, scapula, elbow, forearm, wrist, hand, thumb, head, neck, chest, abdomen, and lower back
Plane of motion	Enumerated	Flexion, extension, abduction, adduction, internal rotation, external rotation, lateral flexion, horizontal abduction, horizontal adduction, protraction, retraction, elevation, depression, inversion, eversion, pronation, supination, plantarflexion, dorsiflexion, radial deviation, ulnar deviation, upward rotation, downward rotation, opposition, forward, backward, lateral, medial, scaption, rotation, closure, clockwise, counterclockwise, distraction, all planes, anterior, posterior, horizontal, vertical, diagonal, and gravity elimination
Duration (seconds)	Integer	N/A[b]
Number of sets	Integer	N/A
Number of reps	Integer	N/A
Exercise purpose	Enumerated	Strength, fine motor, motor control, perception, simulated, power, endurance, joint mobility, joint alignment, muscle flexibility, cardio, pulmonary, agility, and vestibular
Exercise type	Enumerated	Upper extremity strength, lower extremity strength, trunk or core strength, scapular strength, ROM, flexibility or mobility, balance or vestibular, gait training, cardio or aerobic, and functional mobility
Body position	Binary	Weight bearing and non-weight bearing
Negation or hypothetical	Binary	Held or not performed and home exercise program

[a]ROM: range of motion.

[b]N/A: not applicable.

Preprocessing and Section Extraction

Physical therapeutic procedures were usually documented in the section "THERAPY." Therefore, we first filtered out the notes that did not contain a physical therapy visit by excluding files whose names lacked the string "THERAPY." From the resulting set of files, the section on therapeutic procedures was extracted using a regular expression, if such a section existed. This resulted in a total of 23,724 notes, some of which were empty or lacked pertinent information.

The method of section extraction has a few minor limitations. Because the regular expression used to locate these sections assumes a structure in the notes that is not always present, it is possible that a file may contain additional text from other sections of the original note in rare instances. All sections used in the creation of the gold-standard labels were manually examined to ensure the absence of these errors. It is also possible that some therapeutic procedures' sections are completely omitted from the note due to copy-and-paste errors made by their authors.

Because many of the extracted sections were very brief or lacked relevant information, we developed a method to create a more robust set of sections by extracting keywords. Initially, concepts were organized into 9 categories based on the clinical ontology. Each category was then assigned a list of keywords. A section was considered to mention a category if it contained at least 1 of the keywords. Consequently, each section was assigned a score between 0 and 9 based on the number of categories mentioned. All sections with a score of 9 and a random selection of notes with a score of 8 were extracted to generate 300 enriched sections that were anticipated to be relatively dense in information. In addition, 300 random sections were selected,

excluding those with a length of fewer than 200 characters in order to reduce the likelihood of omissions.

Gold-Standard Data Set Creation

Gold standard labels were developed by 2 clinical experts in the field of physical therapy (PD and BA) under the supervision of a senior clinical expert in physical therapy (AB). Each annotator was given a set of guidelines on how to label sections and was told to refer to the clinical ontology for examples of each concept to label. Instructions were given to label explicit mentions of each concept, and inferences were only to be made when specified. For example, the concepts under the categories exercise type and positioning were each given several common keywords that indicate exercises that relate to them. The annotators were given identical batches of 20 randomly selected sections to annotate, and the interannotator agreement was calculated using Fleiss κ. This process was repeated for a total of 3 batches, after which all 3 annotators achieved an interannotator agreement greater than 0.7. Throughout this process, the annotation guidelines were revised, and the structure of the labels was finalized. Once sufficient agreement was reached, 50 sections from the enriched set and 50 more from the random set were given to each annotator, totaling 300 distinct annotated sections. These sections were then split randomly into a training set consisting of 125 sections from each of the original sets and a test set consisting of the remaining 50 sections. The details of this corpus are included in Textbox 1, which outlines the total word count, the number of distinct words, and 2 examples of the data.

Textbox 1. Summary of the annotated corpus.

Total words: 74,104

Total distinct words: 2371

Deidentified note example 1:

- "1: AROM right elbow flx/ext HEP (right arm supported on table) 2: AROM right wrist flx/ext HEP 3: AROM right forearm pronation/supination HEP 4: Thumb opposition HEP 5: Seated AAROM table slide??"

Deidentified note example 2:

- "1: foam balance (heel/toe rocking): x 30 2: step taps with 2 taps from foam 12""" block: x 20 B/L 3: tandem walking: 25' x 2 4: backward walking: 25' x 2 5: foam Lunges: x 20 B/L 6: Dips 4""" stair: 2x10 B/L 7: side stepping green TB 10 ft x5 each direction 9: bridging with LLE leg lift 1""" off mat x10 10: tandem stance on foam x 1' 11: Nustep: L5 x 10' (LEs only)"

Rule-Based NLP

The first NLP method we developed was a named entity recognition (NER) algorithm using MedTagger (OHNLP), which is a software that uses rule-based methods to segment documents and extract named entity information with regular expressions [14]. We used this tool to detect the categories outlined in the ontology by creating explainable rules to extract the physical rehabilitation exercise information and compare it against the gold-standard labels. For each rule defined in the algorithm, MedTagger identified spans of text that matched the expression as well as the corresponding category and concept predicted for that text. We initiated the rules using simple keywords in the clinical ontology as defined in Table 2 and then refined the rules using the training set of the gold-standard notes.

Machine Learning–Based NLP

In addition to attempting to automate the annotation of clinical notes with exercise information, several sequence-level binary classification methods were explored to predict whether a specific concept is mentioned in a given span of text at least once according to the gold-standard labels. Here, a sequence is defined as a string of text within a section that describes an individual exercise. As the therapeutic procedures are documented as numbered lists, it is assumed that each enumerated item that contains text constitutes a single procedure for the purpose of this study. The aim was to extract these procedures from sections and then classify each according to which concepts they mention.

For this task, the sequences provided in the gold-standard data were used as raw input, and targets were defined using the labels that were associated with each sequence. These labels consisted of 101 concepts as given by the clinical ontology in Table 2, excluding duration, sets, and reps since these are numeric types unfit for binary classification tasks. Because the postprocessed output from MedTagger was formatted in a similar manner to the gold-standard data for ease of comparison, a similar method was used to create predictions and directly score MedTagger against the true labels for this task. In this manner, we compared our rule-based NLP algorithm against several other methods by redefining the information extraction task as a sequence classification task. The labels of all predicted spans of text were assigned to the section containing it.

A total of 4 machine learning models were trained to perform binary classification on sections, including SVM [15], LR [16], gradient boosting [17], and AdaBoost [18]. We built different machine learning models for different physical rehabilitation exercise concept extraction tasks. This resulted in 101 distinct SVM, LR, gradient boosting, and AdaBoost models each trained to predict a distinct concept. Each model was created using the *scikit-learn* [19] library in Python (version 3; Python Software Foundation). The input for each model was given in a simple uncased bag-of-words vector space fitted to the training set. The LR was performed with a learning rate of 1×10^{-4} and balanced class weights. The SVM model used a polynomial kernel with a degree of 2 and also used balanced class weights. The AdaBoost and gradient boosting were performed with the default parameters provided by *scikit-learn*, with 100 and 50

estimators, respectively. All unspecified hyperparameters were kept at the default values used by *scikit-learn*.

LLM-Based NLP

Recently, LLMs have gained much interest due to their promising results across many NLP tasks and straightforward development pipelines. To measure a baseline for the performance of LLMs on this data set, this study used OpenAI's ChatGPT, specifically GPT-3.5-turbo [20], which serves as the NLP backend for the first release of ChatGPT. This study accessed ChatGPT through the Microsoft Azure secure cloud service with an approved institutional review board.

For each of the concepts to be identified, a separate prompt was created asking the model to identify whether or not that concept is mentioned in the given sections. We note that the sections used in this exercise were fully deidentified and contained no sensitive information. One of the primary advantages of using LLMs to perform NLP tasks is that they can yield results comparable to other methods with much less time spent on fine-tuning and other manual design tasks. Therefore, in order to evaluate the practicality of using this method in an application with constraints on time or available machine learning expertise, prompts were automatically generated using a template and the name assigned to each label. Two simple prompting techniques were compared: a few-shot prompting technique that included example dialogue for 2 positive examples and 2 negative examples [21] and a zero-shot prompt that omitted any examples [22]. Prompts were created by providing a list of dialogue strings that can be attributed to the user providing the prompt, the model giving the response, and a third party giving instructions to the model, hereafter named System. The following is an example of a few-shot prompt that was automatically generated and sent to ChatGPT.

> *System: You are an assistant assigned to determine if a given text segment from a medical record contains mentions of Active Range of Motion. You must answer yes or no.*
>
> *User: 7: [PERSONALNAME] maze - AROM LUE - 3 rep*
>
> *Model: Yes.*
>
> *User: 23: / Supine on mat / dowel rod no wt above head for triceps*
>
> *Model: No.*
>
> *User: 2: AROM Right wrist ext/flex - 20 x*
>
> *Model: Yes.*
>
> *User: 24: SL HS activation- with tapping and suspended to limit friction x10 DEFERRED*
>
> *Model: No.*
>
> *User: 14: SLR- 2x10 deferred to HEP*

Prompts were made in this manner using 4 dialogue exchanges, switching out the last line for the section to be classified. Any concepts that were labeled fewer than 2 times were excluded from the experiment because not enough examples could be generated for the few-shot prompt. A total of 3 additional concepts were excluded relating to whether a procedure was performed in-office, at home, or not at all, as well as the 3 numeric concepts. This left a total of 82 concepts to be extracted.

Results

Because the clinical ontology covered a broad range of concepts within this domain, many of these concepts were very sparse in our data. More than half of the 101 concepts were present in fewer than 10 exercise descriptions in the train or test sets; these concepts have been omitted from the results. Table 3 contains a breakdown of the F_1-scores for each machine learning method, as well as the performance of the rule-based NLP algorithm on the NER task, for each of the remaining 40 concepts. See Multimedia Appendix 1 for the results on all concepts. The best-performing machine learning model is shown in bold for each concept.

Table 3. Binary F_1-scores of each algorithm on the test set (50 documents).

Category and concept	RBNLP[a] NER[b], n	RBNLP sequence, n	LR[c], n	SVM[d], n	AdaBoost, n	Gradient boosting, n	ChatGPT (few-shot), n	ChatGPT (zero-shot), n	Training set size, n	Test set size, n
Description										
Performed in-office	0.957	0.976	0.970	0.960	0.977	0.983[e]	N/A[f]	N/A	2464	497
Home exercise program	0.986[e]	0.986[e]	0.986[e]	0.938	0.986[e]	0.986[e]	N/A	N/A	93	34
Not performed	0.949	0.949	0.923	0.909	0.936	0.950[e]	N/A	N/A	1295	206
ROM[g]										
Active	0.839	0.830	0.824	0.840	0.863[e]	0.863[e]	0.321	0.109	103	22
Active-assisted	0.769	0.769	0.800	0.791	0.837	0.857[e]	0.543	0.210	160	24
Passive	0.952	0.938	0.970[e]	0.903	0.938	0.970[e]	0.552	0.198	121	16
Side										
Right side	0.912	0.975[e]	0.674	0.851	0.628	0.680	0.912	0.878	548	97
Left side	0.912	0.937[e]	0.763	0.823	0.721	0.752	0.823	0.832	462	134
Bilateral	0.772	0.907[e]	0.559	0.474	0.667	0.659	0.706	0.723	260	51
Location										
Upper extremity	0.847	0.939[e]	0.879	0.847	0.901	0.876	0.291	0.241	285	47
Lower extremity	0.955	0.936	0.936	0.930	0.966	0.978[e]	0.378	0.339	223	44
Hip	0.949	0.947	0.973[e]	0.973[e]	0.943	0.972	0.403	0.806	168	36
Knee	0.950	0.950	0.919	0.882	0.974[e]	0.974[e]	0.469	0.434	108	19
Ankle	1.000[e]	1.000[e]	0.923	0.600	1.000[e]	1.000[e]	0.607	0.262	55	14
Shoulder	0.936	0.977[e]	0.952	0.952	0.953	0.953	0.744	0.548	224	44
Scapula	0.833[e]	0.833[e]	0.783	0.700	0.833[e]	0.833[e]	0.525	0.607	72	10
Elbow	0.967[e]	0.963	0.963	0.943	0.923	0.923	0.848	0.447	147	26
Forearm	0.815	0.833	0.870	0.952[e]	0.870	0.952[e]	0.151	0.204	86	10
Wrist	0.902[e]	0.898	0.826	0.773	0.875	0.875	0.600	0.314	129	23
Hand	0.951[e]	0.944	0.926	0.848	0.925	0.949	0.438	0.574	243	68
Plane										
Abduction	0.976	0.985[e]	0.971	0.937	0.971	0.971	0.576	0.839	170	33
Anterior	0.545	0.545	0.750[e]	0.667	0.750[e]	0.667	0.221	0.195	22	10
Backward	0.727	0.720	0.688	0.800	0.952[e]	0.846	0.720	0.790	92	11
Extension	0.980	0.980	0.979	0.933	0.989[e]	0.989[e]	0.556	0.684	266	48
External rotation	0.897	0.917[e]	0.870	0.818	0.870	0.870	0.655	0.543	74	11
Flexion	0.956	0.947	0.964[e]	0.955	0.964[e]	0.964[e]	0.757	0.615	327	55
Forward	0.977[e]	0.974	0.857	0.865	0.950	0.900	0.667	0.729	148	19

Category and concept	RBNLP[a] NER[b], n	RBNLP sequence, n	LR[c], n	SVM[d], n	AdaBoost, n	Gradient boosting, n	ChatGPT (few-shot), n	ChatGPT (zero-shot), n	Training set size, n	Test set size, n
Lateral	0.577	0.588	0.786	0.837	0.870[e]	0.851	0.546	0.373	132	23
Supination	0.923[e]	0.917	0.880	0.917	0.917	0.917	0.550	0.480	82	11
Exercise type										
Upper extremity strength	0.913[e]	0.913[e]	0.840	0.791	0.913[e]	0.894	0.272	0.166	138	21
Lower extremity strength	0.926	0.969[e]	0.913	0.894	0.924	0.894	0.449	0.332	447	97
Trunk or core strength	0.897	0.889[e]	0.692	0.471	0.471	0.700	0.104	0.090	35	12
Range of motion	0.853	0.876[e]	0.842	0.843	0.725	0.674	0.301	0.153	257	53
Flexibility or mobility	0.962	0.974[e]	0.909	0.857	0.947	0.949	0.279	0.147	178	38
Balance or vestibular	0.787	0.752	0.852	0.809	0.882	0.939[e]	0.597	0.470	351	47
Gait training	0.808	0.837	0.837	0.814	0.851	0.860[e]	0.626	0.529	310	47
Functional mobility	0.775	0.831[e]	0.727	0.750	0.691	0.780	0.220	0.182	204	33
Purpose										
Simulated	0.769	0.769	0.870[e]	0.762	0.857	0.870[e]	0.688	0.667	48	10
Positioning										
Weight bearing	0.788	0.833	0.876[e]	0.867	0.857	0.871	0.197	0.282	255	43
Non-weight bearing	0.931	0.932	0.916	0.918	0.946[e]	0.923	0.283	0.038	539	91
Average	0.878	0.891[e]	0.861	0.835	0.875	0.883	0.502	0.433	283	53

[a]RBNLP: rule-based natural language processing.

[b]NER: named entity recognition.

[c]LR: logistic regression.

[d]SVM: support vector machine.

[e]The best performance for each entity.

[f]N/A: not applicable.

[g]ROM: range of motion.

The rule-based NLP's performance on the sequence classification task was similar to its performance on the NER task. Instances of higher performance in sequence classification compared to NER can be partially explained by mismatches in predicted spans and their labels affecting NER accuracy, yet still allowing for correct overall text section classification. The rule-based algorithm tied with or outperformed the other models on half of the concepts in Table 3. Among the machine learning models, gradient boosting performed nearly as well, achieving the highest F_1-score on 18 concepts.

In addition to these concepts, the rule-based NLP algorithm also predicted the spans of durations, sets, and reps. Since these categories do not have any specific concepts assigned to them, the number presented in each span was used instead as a comparison against the true label, converting minutes to seconds where applicable. This resulted in F_1-scores of 0.65, 0.58, and 0.88, respectively. It is important to note that we limited the experiments for "Duration," "Sets," and "Reps" exclusively to rule-based algorithms because these categories inherently involve numeric data, which align well with the deterministic and pattern-based nature of rule-based approaches.

Gradient boosting demonstrated the best performance for identifying range of motion (ROM) concepts and determining the location of exercise (performed in-office, home exercise program, and not performed) with F_1-scores of 0.863 for active ROM; 0.857 for active-assisted ROM; and 0.977, 0.986, and 0.950, respectively, for the locations. The rule-based natural language processing algorithm outperformed machine learning models in detecting sides of the body with F_1-scores of 0.975 for the right side and 0.937 for the left side, and it also performed the best on most exercise types, except for balance or vestibular and gait training concepts, which were classified best by gradient boosting with F_1-scores of 0.939 and 0.860, respectively. The LR obtained a strictly higher score than other methods in the weight-bearing exercise concept with an F_1-score of 0.876. The AdaBoost got a strictly higher score on 3 concepts, notably on non–weight bearing positioning with an F_1-score of 0.946. The SVM model did not score higher than other models but had 3 ties, indicating competitive performance.

These findings indicate that the rule-based approach is particularly effective for certain types of exercises, with superior performance in most categories. However, gradient boosting demonstrated strength in more complex categorizations such as balance or vestibular and gait training, where understanding nuanced differences is crucial.

For the LLM-based NLP, the results show that both zero-shot prompts and few-shot prompts result in high recall scores that sometimes exceed other methods. However, precision was quite low for most concepts, and F_1-scores did not exceed every other method for any concept. However, ChatGPT did occasionally outperform some of the simpler machine learning models and, on 1 occasion, even outperformed the rule-based algorithm (on the backward plane of motion concept). The average precision over all 82 concepts tested was 0.33 for the zero-shot approach and 0.27 for the few-shot approach. The average recall was 0.8 for the zero-shot approach and 0.82 for the few-shot approach. This resulted in average F_1-scores of 0.37 and 0.35, respectively, indicating that the zero-shot approach was slightly better on average than the few-shot approach. However, the few-shot approach performed the best for all but 10 concepts. The reason the zero-shot method performed better on average is thus due to the fact that it shows significant improvement on a few specific concepts, such as hip, scapula, hand, abduction, and extension.

Discussion

Observations

As indicated by the high performance of the machine learning models on many of the concepts, the task of extracting information from exercise descriptions was not complex. Although some of these concepts could be extracted effectively using straightforward rules or a small machine learning model, there were also many cases where clinical notes appeared ambiguous without context. For instance, the abbreviation "SL" could be interpreted as "single leg" or "side-lying" depending on the exercise being described. In addition, "L" could mean "left" or "lateral," which explains why the rule-based NLP algorithm performed slightly worse when classifying left versus right. The use of single letters as abbreviations, especially "A" as a shorthand for "anterior," could cause issues in machine learning algorithms without careful consideration. It would be possible to increase the performance of the rule-based algorithm by further tuning the rules to search for context clues at other points in the document, but this could potentially cause the rules to overfit the training set. Of particular interest are the numeric data present in duration, sets, and reps. These are particularly tricky to extract since they are expressed in a wide variety of ways by different physicians. It can be difficult to define what sets and reps are depending on the exercise, and sometimes one or both are not well-defined at all. Additionally, the use of apostrophes and quotes can either indicate measurements of time or distance, once again requiring context to disambiguate. Mentions of distance were not annotated in the gold-standard labels, but it is important in measuring the intensity of some exercises, so we plan to include it in the future.

Some of the misclassifications of the rule-based algorithm are due to inaccuracies in the gold-standard data set. For instance, many false positives produced by the rule-based algorithm appeared to be concepts that were missed by the annotators. There were also a few minor errors that could be explained by a mouse slip, including a span of text being assigned the wrong concept or a span excluding the last letter in a word. There were also some spelling mistakes in the notes themselves; common instances were explicitly mentioned in the rules to increase precision. Preprocessing clinical notes to correct spelling mistakes might be useful to improve results, although this creates a risk of incorrect changes being made to uncommon words. All of these errors were not particularly common throughout the labels, but they could have a significant effect on concepts that are already uncommon in the data.

Another obstacle that obscured some of the signals in the data came from the deidentification process. In addition to removing names, addresses, and other protected information from these documents, many other tokens and phrases were mistakenly removed, including equipment names and numbers denoting indices in a list. These were replaced with placeholder tokens such as "[ADDRESS]" or "[PERSONALNAME]." The low precision of the deidentification process caused some relevant information to be obfuscated or entirely erased from notes.

During the data annotation, we found that many of the concepts identified as relevant in this domain were not well documented in the data we extracted for annotation. This could be due in part to the fact that the data were only collected from patients who had a stroke, but this is not expected to be the main reason because patients who had a stroke can have a wide variety of musculoskeletal problems, resulting in a correspondingly wide variety of treatments being mentioned in clinical notes [23]. The other reason the data set lacks many of these concepts could be that they are rarely mentioned in these particular sections of clinical notes, either because they are not common enough to appear in many records at all or because they are mentioned more often in other sections. Thus, future research could focus on improving extraction methods to focus more on these uncommon concepts or include information from outside of the exercise descriptions.

In addition to ChatGPT for the LLM-based NLP approach, we also fine-tuned a Bidirectional Encoder Representations from Transformers (BERT) model with the task of categorizing the physical rehabilitation exercise concept. The BioClinicalBERT model [24] was used, which was pretrained on Medical Information Mart for Intensive Care-III (MIMIC-III) [25]. However, the amount of data collected seemed insufficient to make the model perform comparably to simpler methods. The model with the highest F_1-score on the validation set had an average F_1-score of 0.05 across all concepts on the test set. It accurately predicted in-office exercise performance with an F_1-score of 0.72. However, the performance on the remaining 100 concepts ranged only from 0 to 0.35. Therefore, we did not include this approach in the experimental comparison.

Limitations and Future Work

One limitation in this research was the necessary exclusion of "Duration," "Number of Sets," and "Number of Reps" from our machine learning–based NLP analysis due to their numeric nature, rendering them unsuitable for binary classification tasks. In future work, we plan to incorporate regression models or specialized classification techniques capable of handling numeric data. We also plan to expand our research to include additional variables such as stroke duration and severity, recognizing their potential to significantly enhance the prediction accuracy and effectiveness of rehabilitation strategies.

Furthermore, another limitation of this study is that we did not consider technique names and their association with specific motion types in rehabilitation exercise notes. For instance, we encountered the text "1: Standing AAROM PNF exercise D1/D2 flexion - 20 x" during annotation but did not annotate the technique name PNF (proprioceptive neuromuscular facilitation). To address this, in future work, we intend to develop a supplementary module for our algorithm that can effectively extract and map popular technique names to their corresponding motion types and categories, thereby enhancing the comprehensiveness and applicability of the algorithm.

Moreover, we plan to implement a robust standardized extraction protocol in the next version of our algorithm to mitigate the omission of therapeutic procedure sections due to copy-and-paste errors. This protocol will include multiple checks for consistency and completeness and will be assessed through a pilot study to ensure its reliability and accuracy. To enhance our model's generalizability amid varied note-writing practices across rehabilitation facilities, future research will also focus on diversifying data sources, refining adaptability to diverse writing styles and terminologies, and conducting extensive validation studies in a range of settings to improve performance. Through continuous monitoring and refinement of our extraction process, we are committed to enhancing the reliability and validity of our data, thereby strengthening the overall quality and impact of our research.

Conclusions

In this study, we developed and evaluated several NLP algorithms to extract physical rehabilitation exercise information from clinical notes of patients who had stroke. We first created a novel and comprehensive clinical ontology to represent physical rehabilitation exercise in clinical notes and then developed a variety of NLP algorithms leveraging state-of-the-art techniques, including rule-based NLP algorithms, machine learning–based NLP algorithms, and LLM-based NLP algorithms. The experiments on the clinical notes of a cohort of patients who had a stroke showed that the rule-based NLP algorithm had the best performance for most of the physical rehabilitation exercise concepts. Among all machine learning models, gradient boosting achieved the best performance on a majority of concepts. On the other hand, the rule-based NLP performed well for extracting handled durations, sets, and reps, while gradient boosting excelled in ROM and location detection. The LLM-based NLP achieved high recall with zero-shot and few-shot prompts but low precision and F_1-scores. It occasionally outperformed simpler models and once bet the rule-based algorithm.

Acknowledgments

This work was supported by the School of Health and Rehabilitation Sciences Dean's Research and Development Award.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full binary F_1-scores of each algorithm on the test set (50 documents) and additional results from the ChatGPT experiment.
[DOCX File , 63 KB - medinform_v12i1e52289_app1.docx]

References

1. Ginsburg GS, Phillips KA. Precision medicine: from science to value. Health Aff (Millwood) 2018;37(5):694-701 [FREE Full text] [doi: 10.1377/hlthaff.2017.1624] [Medline: 29733705]
2. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci 2021;14(1):86-93 [FREE Full text] [doi: 10.1111/cts.12884] [Medline: 32961010]
3. Shin SH, Bode AM, Dong Z. Precision medicine: the foundation of future cancer therapeutics. NPJ Precis Oncol 2017;1(1):12 [FREE Full text] [doi: 10.1038/s41698-017-0016-z] [Medline: 29872700]

4. French MA, Roemmich RT, Daley K, Beier M, Penttinen S, Raghavan P, et al. Precision rehabilitation: optimizing function, adding value to health care. Arch Phys Med Rehabil 2022;103(6):1233-1239. [doi: 10.1016/j.apmr.2022.01.154] [Medline: 35181267]
5. Severin R, Sabbahi A, Arena R, Phillips SA. Precision medicine and physical therapy: a healthy living medicine approach for the next century. Phys Ther 2022;102(1):pzab253 [FREE Full text] [doi: 10.1093/ptj/pzab253] [Medline: 34718788]
6. Lotze M, Moseley GL. Theoretical considerations for chronic pain rehabilitation. Phys Ther 2015;95(9):1316-1320 [FREE Full text] [doi: 10.2522/ptj.20140581] [Medline: 25882484]
7. Blum C, Baur D, Achauer LC, Berens P, Biergans S, Erb M, et al. Personalized neurorehabilitative precision medicine: from data to therapies (MWKNeuroReha)—a multi-centre prospective observational clinical trial to predict long-term outcome of patients with acute motor stroke. BMC Neurol 2022;22(1):238 [FREE Full text] [doi: 10.1186/s12883-022-02759-2] [Medline: 35773640]
8. 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;23(3):e22951 [FREE Full text] [doi: 10.2196/22951] [Medline: 33683212]
9. Newman-Griffis D, Maldonado JC, Ho PS, Sacco M, Silva RJ, Porcino J, et al. Linking free text documentation of functioning and disability to the ICF with natural language processing. Front Rehabil Sci 2021;2:742702 [FREE Full text] [doi: 10.3389/fresc.2021.742702] [Medline: 35694445]
10. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-03-18]
11. International Classification of Functioning, Disability, and Health Children and Youth Version: ICF-CY. Geneva: World Health Organization; 2007.
12. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279-290. [Medline: 17095826]
13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
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. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273-297 [FREE Full text] [doi: 10.1007/bf00994018]
16. Pregibon D. Logistic regression diagnostics. Ann Statist 1981;9(4):705-724 [FREE Full text] [doi: 10.1214/aos/1176345513]
17. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001;29(5):1189-1232 [FREE Full text] [doi: 10.1214/aos/1013203451]
18. Schapire RE. Explaining adaboost. In: Schölkopf B, Vovk V, Luo Z, editors. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Berlin Heidelberg: Springer; 2013:37-52.
19. 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]
20. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Red Hook, New York, US: Curran Associates, Inc; 2022:27730-27744.
21. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. ArXiv 2023 [FREE Full text]
22. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981 [FREE Full text] [Medline: 37128372]
23. De Rosario H, Pitarch-Corresa S, Pedrosa I, Vidal-Pedrós M, de Otto-López B, García-Mieres H, et al. Applications of natural language processing for the management of stroke disorders: scoping review. JMIR Med Inform 2023;11:e48693 [FREE Full text] [doi: 10.2196/48693] [Medline: 37672328]
24. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. ArXiv 2019 [FREE Full text] [doi: 10.48550/arXiv.1904.03323]
25. 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]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
ICF: International Classification of Functioning, Disability, and Health
LLM: large language model
LR: logistic regression
MIMIC-III: Medical Information Mart for Intensive Care-III
NER: named entity recognition
NLP: natural language processing

PNF: proprioceptive neuromuscular facilitation
ROM: range of motion
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
SVM: support vector machine
UMLS: Unified Medical Language System
UPMC: University of Pittsburgh Medical Center

Edited by A Benis; submitted 29.08.23; peer-reviewed by Z Alhassan, A Rehan Youssef; comments to author 27.11.23; revised version received 02.01.24; accepted 27.02.24; published 03.04.24.

Please cite as:
Sivarajkumar S, Gao F, Denny P, Aldhahwani B, Visweswaran S, Bove A, Wang Y
Mining Clinical Notes for Physical Rehabilitation Exercise Information: Natural Language Processing Algorithm Development and Validation Study
JMIR Med Inform 2024;12:e52289
URL: https://medinform.jmir.org/2024/1/e52289
doi:10.2196/52289
PMID:38568736

Original Paper

An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study

Sonish Sivarajkumar[1], BS; Mark Kelley[2], MS; Alyssa Samolyk-Mazzanti[2], MS; Shyam Visweswaran[1,3], MD, PhD; Yanshan Wang[1,2,3], PhD

[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, United States

[2]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA, United States

[3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
Yanshan Wang, PhD
Department of Health Information Management
University of Pittsburgh
6026 Forbes Tower
Pittsburgh, PA, 15260
United States
Phone: 1 4123832712
Email: yanshan.wang@pitt.edu

Abstract

Background: Large language models (LLMs) have shown remarkable capabilities in natural language processing (NLP), especially in domains where labeled data are scarce or expensive, such as the clinical domain. However, to unlock the clinical knowledge hidden in these LLMs, we need to design effective prompts that can guide them to perform specific clinical NLP tasks without any task-specific training data. This is known as in-context learning, which is an art and science that requires understanding the strengths and weaknesses of different LLMs and prompt engineering approaches.

Objective: The objective of this study is to assess the effectiveness of various prompt engineering techniques, including 2 newly introduced types—heuristic and ensemble prompts, for zero-shot and few-shot clinical information extraction using pretrained language models.

Methods: This comprehensive experimental study evaluated different prompt types (simple prefix, simple cloze, chain of thought, anticipatory, heuristic, and ensemble) across 5 clinical NLP tasks: clinical sense disambiguation, biomedical evidence extraction, coreference resolution, medication status extraction, and medication attribute extraction. The performance of these prompts was assessed using 3 state-of-the-art language models: GPT-3.5 (OpenAI), Gemini (Google), and LLaMA-2 (Meta). The study contrasted zero-shot with few-shot prompting and explored the effectiveness of ensemble approaches.

Results: The study revealed that task-specific prompt tailoring is vital for the high performance of LLMs for zero-shot clinical NLP. In clinical sense disambiguation, GPT-3.5 achieved an accuracy of 0.96 with heuristic prompts and 0.94 in biomedical evidence extraction. Heuristic prompts, alongside chain of thought prompts, were highly effective across tasks. Few-shot prompting improved performance in complex scenarios, and ensemble approaches capitalized on multiple prompt strengths. GPT-3.5 consistently outperformed Gemini and LLaMA-2 across tasks and prompt types.

Conclusions: This study provides a rigorous evaluation of prompt engineering methodologies and introduces innovative techniques for clinical information extraction, demonstrating the potential of in-context learning in the clinical domain. These findings offer clear guidelines for future prompt-based clinical NLP research, facilitating engagement by non-NLP experts in clinical NLP advancements. To the best of our knowledge, this is one of the first works on the empirical evaluation of different prompt engineering approaches for clinical NLP in this era of generative artificial intelligence, and we hope that it will inspire and inform future research in this area.

(JMIR Med Inform 2024;12:e55318) doi:10.2196/55318

KEYWORDS

large language model; LLM; LLMs; natural language processing; NLP; in-context learning; prompt engineering; evaluation; zero-shot; few shot; prompting; GPT; language model; language; models; machine learning; clinical data; clinical information; extraction; BARD; Gemini; LLaMA-2; heuristic; prompt; prompts; ensemble

Introduction

Clinical information extraction (IE) is the task of identifying and extracting relevant information from clinical narratives, such as clinical notes, radiology reports, or pathology reports. Clinical IE has many applications in health care, such as improving diagnosis, treatment, and decision-making; facilitating clinical research; and enhancing patient care [1,2]. However, clinical IE faces several challenges, such as the scarcity and heterogeneity of annotated data, the complexity and variability of clinical language, and the need for domain knowledge and expertise.

Zero-shot IE is a promising paradigm that aims to overcome these challenges by leveraging large pretrained language models (LMs) that can perform IE tasks without any task-specific training data [3]. In-context learning is a framework for zero-shot and few-shot learning, where a large pretrained LM takes a context and directly decodes the output without any retraining or fine-tuning [4]. In-context learning relies on prompt engineering, which is the process of crafting informative and contextually relevant instructions or queries as inputs to LMs to guide their output for specific tasks [5]. The use of prompt engineering lies in its ability to leverage the powerful capabilities of large LMs (LLMs), such as GPT-3.5 (OpenAI) [6], Gemini (Google) [7], LLaMA-2 (Meta) [8], even in scenarios where limited or no task-specific training data are available. In clinical natural language processing (NLP), where labeled data sets tend to be scarce, expensive, and time-consuming to create, splintered across institutions, and constrained by data use agreements, prompt engineering becomes even more crucial to unlock the potential of state-of-the-art LLMs for clinical NLP tasks.

While prompt engineering has been widely explored for general NLP tasks, its application and impact in clinical NLP remain relatively unexplored. Most of the existing literature on prompt engineering in the health care domain focuses on biomedical NLP tasks rather than clinical NLP tasks that involve processing real-world clinical notes. For instance, Chen et al [9] used a fixed template as the prompt to measure the performance of LLMs on biomedical NLP tasks but did not investigate different kinds of prompting methods. Wang et al [10] gave a comprehensive survey of prompt engineering for health care NLP applications such as question-answering systems, text summarization, and machine translation. However, they did not compare and evaluate different types of prompts for specific clinical NLP tasks and how the performance varies across different LLMs. There is a lack of systematic and comprehensive studies on how to engineer prompts for clinical NLP tasks, and the existing literature predominantly focuses on general NLP problems. This creates a notable gap in the research, warranting a dedicated investigation into the design and development of effective prompts specifically for clinical NLP. Currently, researchers in the field lack a comprehensive understanding of the types of prompts that exist, their relative effectiveness, and the challenges associated with their implementation in clinical settings.

The main research question and objectives of this study are to investigate how to engineer prompts for clinical NLP tasks, identify best practices, and address the challenges in this emerging field. By doing so, we aim to propose a guideline for future prompt-based clinical NLP studies. In this work, we present a comprehensive empirical evaluation study on prompt engineering for 5 diverse clinical NLP tasks, namely, clinical sense disambiguation, biomedical evidence extraction, coreference resolution, medication status extraction, and medication attribute extraction [11,12]. By systematically evaluating different types of prompts proposed in recent literature, including prefix [13], cloze [14], chain of thought [15], and anticipatory prompts [16], we gain insights into their performance and suitability for each task. Two new types of prompting approaches were also introduced: (1) heuristic prompts and (2) ensemble prompts. The rationale behind these novel prompts is to leverage the existing knowledge and expertise in rule-based NLP, which has been prominent and has shown significant results in the clinical domain [17]. We hypothesize that heuristic prompts, which are based on rules derived from domain knowledge and linguistic patterns, can capture the salient features and constraints of the clinical IE tasks. We also conjecture that ensemble prompts, which are composed of multiple types of prompts, can benefit from the complementary strengths and mitigate the weaknesses of each individual prompt.

One of the key aspects of prompt engineering is the number of examples or shots that are provided to the model along with the prompt. Few-shot prompting is a technique that provides the model with a few examples of input-output pairs, while zero-shot prompting does not provide any examples [3,18]. By contrasting these strategies, we aim to shed light on the most efficient and effective ways to leverage prompt engineering in clinical NLP. Finally, we propose a prompt engineering framework to build and deploy zero-shot NLP models for the clinical domain. This study covers 3 state-of-the-art LMs, including GPT-3.5, Gemini, and LLaMA-2, to assess the generalizability of the findings across various models. This work yields novel insights and guidelines for prompt engineering specifically for clinical NLP tasks.

Methods

Tasks

We selected 5 distinct clinical NLP tasks representing diverse categories of natural language understanding: clinical sense disambiguation (text classification) [19], biomedical evidence extraction (named entity recognition) [20], coreference resolution [21], medication status extraction (named entity recognition+classification) [22], and medication attribute

extraction (named entity recognition+relation extraction) [23]. Table 1 provides a succinct overview of each task, an example scenario, and the corresponding prompt type used for each task.

Table 1. Task descriptions.

Task	NLP[a] task category	Description	Example prompt
Clinical sense disambiguation	Text classification	This task involves identifying the correct meaning of clinical abbreviations within a given context.	What is the meaning of the abbreviation CR[b] in the context of cardiology?
Biomedical evidence extraction	Text extraction	In this task, interventions are extracted from biomedical abstracts.	Identify the psychological interventions in the given text?
Coreference resolution	Coreference resolution	The goal here is to identify all mentions in clinical text that refer to the same entity.	Identify the antecedent for the patient in the clinical note.
Medication status extraction	NER[c]+classification	This task involves identifying whether a medication is currently being taken, not taken, or unknown.	What is the current status of [24] in the treatment of [25]?
Medication attribute extraction	NER+RE[d]	The objective here is to identify specific attributes of a medication, such as dosage and frequency.	What is the recommended dosage of [26] for [27] and how often?

[a]NLP: natural language processing.

[b]CR: cardiac resuscitation.

[c]NER: named entity recognition.

[d]RE: relation extraction.

Data Sets and Evaluation

The prompts were evaluated on 3 LLMs, GPT-3.5, Gemini, and LLaMA-2, under both zero-shot and few-shot prompting conditions, using precise experimental settings and parameters. To simplify the evaluation process and facilitate clear comparisons, we adopted accuracy as the sole evaluation metric for all tasks. Accuracy is defined as the proportion of correct outputs generated by the LLM for each task, using a resolver that maps the output to the label space. Table 2 shows the data sets and sample size for each clinical NLP task. The data sets are as follows:

- Clinical abbreviation sense inventories: This is a data set of clinical abbreviations, senses, and instances [28]. It contains 41 acronyms from 18,164 notes, along with their expanded forms and contexts. We used a randomly sampled subset from this data set for clinical sense disambiguation, coreference resolution, medication status extraction, and medication attribute extraction tasks (Table 2).
- Evidence-based medicine-NLP: This is a data set of evidence-based medicine annotations for NLP [29]. It contains 187 abstracts and 20 annotated abstracts, with interventions extracted from the text. We used this data set for the biomedical evidence extraction task.

Table 2. Evaluation data sets and samples for different tasks.

Task	Data set	Data set example	Samples
Clinical sense disambiguation	CASI[a]	The abbreviation "CR[b]" can refer to "cardiac resuscitation" or "computed radiography."	11 acronyms from 55 notes
Biomedical evidence extraction	EBM[c]-NLP[d]	Identifying panic, avoidance, and agoraphobia (psychological interventions)	187 abstracts and 20 annotated abstracts
Coreference resolution	CASI	Resolving references to "the patient" or "the study" within a clinical trial report.	105 annotated examples
Medication status extraction	CASI	Identifying that a patient is currently taking insulin for diabetes.	105 annotated examples with 340 medication status pairs
Medication attribute extraction	CASI	Identifying dosage, frequency, and route of a medication for a patient.	105 annotated examples with 313 medications and 533 attributes

[a]CASI: clinical abbreviation sense inventories.

[b]CR: cardiac resuscitation.

[c]EBM: evidence-based medicine.

[d]NLP: natural language processing.

All experiments were carried out in different system settings. All GPT-3.5 experiments were conducted using the GPT-3.5 Turbo application programming interface as of the September 2023 update. The LLaMA-2 model was directly accessed for

our experiments. Gemini was accessed using the Gemini application (previously BARD)—Google's generative artificial intelligence conversational system. These varied system settings and access methods were taken into account to ensure the reliability and validity of our experimental results, given the differing architectures and capabilities of each LLM.

In evaluating the prompt-based approaches on GPT-3.5, Gemini, and LLaMA-2, we have also incorporated traditional NLP baselines to provide a comprehensive understanding of the LLMs' performance in a broader context. These baselines include well-established models such as Bidirectional Encoder Representations From Transformers (BERT) [30], Embeddings From Language Models (ELMO) [31], and PubMedBERT-Conditional Random Field (PubMedBERT-CRF) [32], which have previously set the standard in clinical NLP tasks. By comparing the outputs of LLMs against these baselines, we aim to offer a clear perspective on the advancements LLMs represent in the field. This comparative analysis is crucial for appreciating the extent to which prompt engineering techniques can leverage the inherent capabilities of LLMs, marking a significant evolution from traditional approaches to more dynamic and contextually aware methodologies in clinical NLP.

Prompt Creation Process

A rigorous process was followed to create suitable prompts for each task. These prompts were carefully crafted to match the specific context and objectives of each task. There is no established method for prompt design and selection as of now. Therefore, we adopted an iterative approach where prompts, which are created by health care experts, go through a verification and improvement process in an iterative cycle, which involved design, experimentation, and evaluation, as depicted in Figure 1.

Figure 1. Iterative prompt design process: a schematic diagram of the iterative prompt creation process for clinical NLP tasks. The process consists of 3 steps: sampling, prompt designing, and deployment. The sampling step involves defining the task and collecting data and annotations. The prompt designing step involves creating and refining prompts using different types and language models. The deployment step involves selecting the best model and deploying the model for clinical use. LLM: large language model; NER: named entity recognition; NLP: natural language processing; RE: relation extraction.

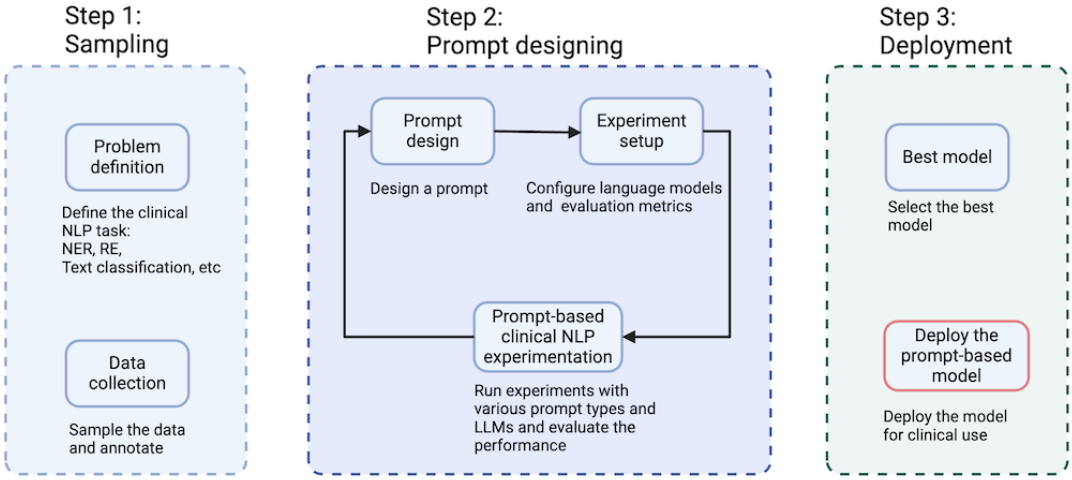


Figure 1 illustrates the 3 main steps of our prompt creation process: sampling, prompt designing, and deployment. In the sampling step (step 1), we defined the clinical NLP task (eg, named entity recognition, relation extraction, and text classification) and collected a sample of data and annotations as an evaluation for the task. In the prompt designing step (step 2), a prompt was designed for the task using one of the prompt types (eg, simple prefix prompt, simple cloze prompt, heuristic prompt, chain of thought prompt, question prompt, and anticipatory prompt). We also optionally performed few-shot prompting by providing some examples along with the prompt. The LLMs and the evaluation metrics for the experiment setup were then configured. We ran experiments with various prompt types and LLMs and evaluated their performance on the task. Based on the results, we refined or modified the prompt design until we achieved satisfactory performance or reached a limit. In the deployment step (step 3), the best prompt-based models were selected based on their performance metrics, and the model was deployed for the corresponding task.

Prompt Engineering Techniques

Overview

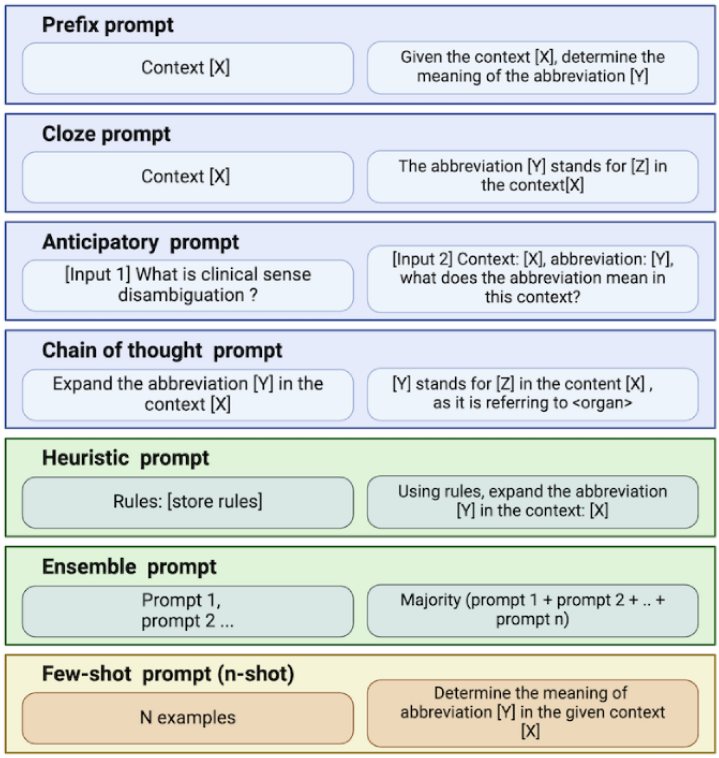
Prompt engineering is the process of designing and creating prompts that elicit desired responses from LLMs. Prompts can be categorized into different types based on their structure, function, and complexity.

Each prompt consists of a natural language query that is designed to elicit a specific response from the pretrained LLM. The prompts are categorized into 7 types, as illustrated in Figure 2 (all prompts have been included in Multimedia Appendix 1). Prefix prompts are the simplest type of prompts, which prepend a word or phrase indicating the type or format or tone of response for control and relevance. Cloze prompts are based on the idea of fill in the blank exercises, which create a masked token in the input text and ask the LLM to predict the missing word or phrase [3]. Anticipatory prompts are the prompts anticipating the next question or command based on experience or knowledge, guiding the conversation. Chain of thought

prompting involves a series of intermediate natural language reasoning steps that lead to the final output [15].

In addition to the existing types of prompts, 2 new novel prompts were also designed: heuristic prompts and ensemble prompts, which will be discussed in the following sections.

Figure 2. Types of prompts: examples of 7 types of prompts that we used to query the pretrained language model for different clinical information extraction tasks. [X]: context; [Y]: abbreviation; [Z]: expanded form.



Heuristic Prompts

Heuristic prompts are rule-based prompts that decompose complex queries into smaller, more manageable components for comprehensive answers. Adopting the principles of traditional rule-based NLP, which relies on manually crafted, rule-based algorithms for specific clinical NLP applications, we have integrated these concepts into our heuristic prompts approach. These prompts use a set of predefined rules to guide the LLM in expanding abbreviations within a given context. For instance, a heuristic prompt might use the rule that an abbreviation is typically capitalized, followed by a period, and preceded by an article or a noun. This approach contrasts with chain of thought prompts, which focus on elucidating the reasoning or logic behind an output. Instead, heuristic prompts leverage a series of predefined rules to direct the LLM in executing a specific task.

Mathematically, we can express a heuristic prompt as $H(x)$, a function applied to an input sequence x. This function is defined as a series of rule-based transformations T_i, where i indicates the specific rule applied. The output of this function, denoted as y_H, is then:

$$y_H=H(x)=T_n(T_{\{n-1\}}(\ldots T_1(x)))$$

Here, each transformation T_i applies a specific heuristic rule to modify the input sequence, making it more suitable for processing by LLMs.

From an algorithmic standpoint, heuristic prompts are implemented by defining a set of rules $R=\{R_1, R_2, \ldots, R_m\}$. Each rule R_j is a function that applies a specific heuristic criterion to an input token or sequence of tokens. Algorithmically, the heuristic prompting process can be summarized as follows:

By merging the precision and specificity of traditional rule-based NLP methods with the advanced capabilities of LLMs, the heuristic prompts offer a robust and accurate system for clinical information processing and analysis.

Ensemble Prompts

Ensemble prompts are prompts that combine multiple prompts using majority voting for aggregated outputs. They use various types of prompts to generate multiple responses to the same input, subsequently selecting the most commonly occurring output as the final answer. For instance, an ensemble prompt might use 3 different prefix prompts, or a combination of other prompt types, to produce 3 potential expansions for an abbreviation. The most frequently appearing expansion is then chosen. For the sake of simplicity, we amalgamated the outputs from all 5 different prompt types using a majority voting approach.

Mathematically, consider a set of m different prompting methods $P_1, P_2, \ldots, P_m$ applied to the same input x. Each method generates

an output y_i for i=1,2, ..., m. The ensemble prompt's output y_E is then the mode of these outputs:

y_E=mode (y_1, y_2, ..., y_m)

Algorithmically, the ensemble prompting process is as follows:

The rationale behind an ensemble prompt is that by integrating multiple types of prompts, we can use the strengths and counterbalance the weaknesses of each individual prompt, offering a robust and potentially more accurate response. Some prompts may be more effective for specific tasks or models, while others might be more resilient to noise or ambiguity. Majority voting allows us to choose the most likely correct or coherent output from the variety generated by different prompt types.

Results

Overview

In this section, we present the results of our experiments on prompt engineering for zero-shot clinical IE. Various prompt types were evaluated across 5 clinical NLP tasks, aiming to understand how different prompts influence the accuracy of different LLMs. Zero-shot and few-shot prompting strategies were also compared, exploring how the addition of context affects the model performance. Furthermore, we tested an ensemble approach that combines the outputs of different prompt types using majority voting. Finally, the impact of different LLMs on task performance was analyzed, and some interesting patterns were observed. Table 3 illustrates that different prompt types have different levels of effectiveness for different tasks and LLMs. We can also observe some general trends across the tasks and models.

Table 3. Performance comparison of different prompt types and language models.

Task and language model	Simple prefix	Simple cloze	Anticipatory	Heuristic	Chain of thought	Ensemble	Few shot
Clinical sense disambiguation							
GPT-3.5	0.88	0.86	0.88	0.96[a]	0.9	0.9	0.82
Gemini	0.76[b]	0.68	0.71	0.75	0.72	0.71	0.67
LLaMA-2	0.88[b]	0.76	0.82	0.82	0.78	0.82	0.78
BERT[c] (from [33])	0.42	0.42	0.42	0.42	0.42	0.42	0.42
ELMO[d] (from [33])	0.55	0.55	0.55	0.55	0.55	0.55	0.55
Biomedical evidence extraction							
GPT-3.5	0.92	0.82	0.88	0.94	0.94	0.88	0.96[a]
Gemini	0.89	0.89	0.91[b]	0.9	0.91[b]	0.9	0.88
LLaMA-2	0.85	0.88[b]	0.87	0.88[b]	0.87	0.88	0.86
PubMedBERT-CRF[e] (from [29])	0.35	0.35	0.35	0.35	0.35	0.35	0.35
Coreference resolution							
GPT-3.5	0.78	0.6	0.74	0.94[a]	0.94[a]	0.74	0.74
Gemini	0.69	0.81[b]	0.73	0.67	0.71	0.69	0.7
LLaMA-2	0.8[b]	0.64	0.74	0.76	0.8[b]	0.78	0.68
Toshniwal et al [34]	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Medication status extraction							
GPT-3.5	0.76[a]	0.72	0.75	0.74	0.73	0.75	0.72
Gemini	0.67[b]	0.51	0.65	0.55	0.59	0.58	0.55
LLaMA-2	0.58	0.48	0.52	0.64[b]	0.52	0.58	0.42
ScispaCy [35]	0.52	0.52	0.52	0.52	0.52	0.52	0.52
Medication attribute extraction							
GPT-3.5	0.88	0.84	0.9	0.96[a]	0.96[a]	0.9	0.96[a]
Gemini	0.68	0.72	0.88[c]	0.7	0.74	0.76	0.88[b]
LLaMA-2	0.6	0.66	0.58	0.66	0.72[b]	0.64	0.6
ScispaCy	0.70	0.70	0.70	0.70	0.70	0.70	0.70

[a]Best performance on a task regardless of the model (ie, for each GPT-3.5 or Gemini or LLaMA-2 triple).

[b]Best performance for each model on a task.

[c]BERT: Bidirectional Encoder Representations From Transformers.

[d]ELMO: Embeddings From Language Models.

[e]PubMedBERT-CRF: PubMedBERT-Conditional Random Field.

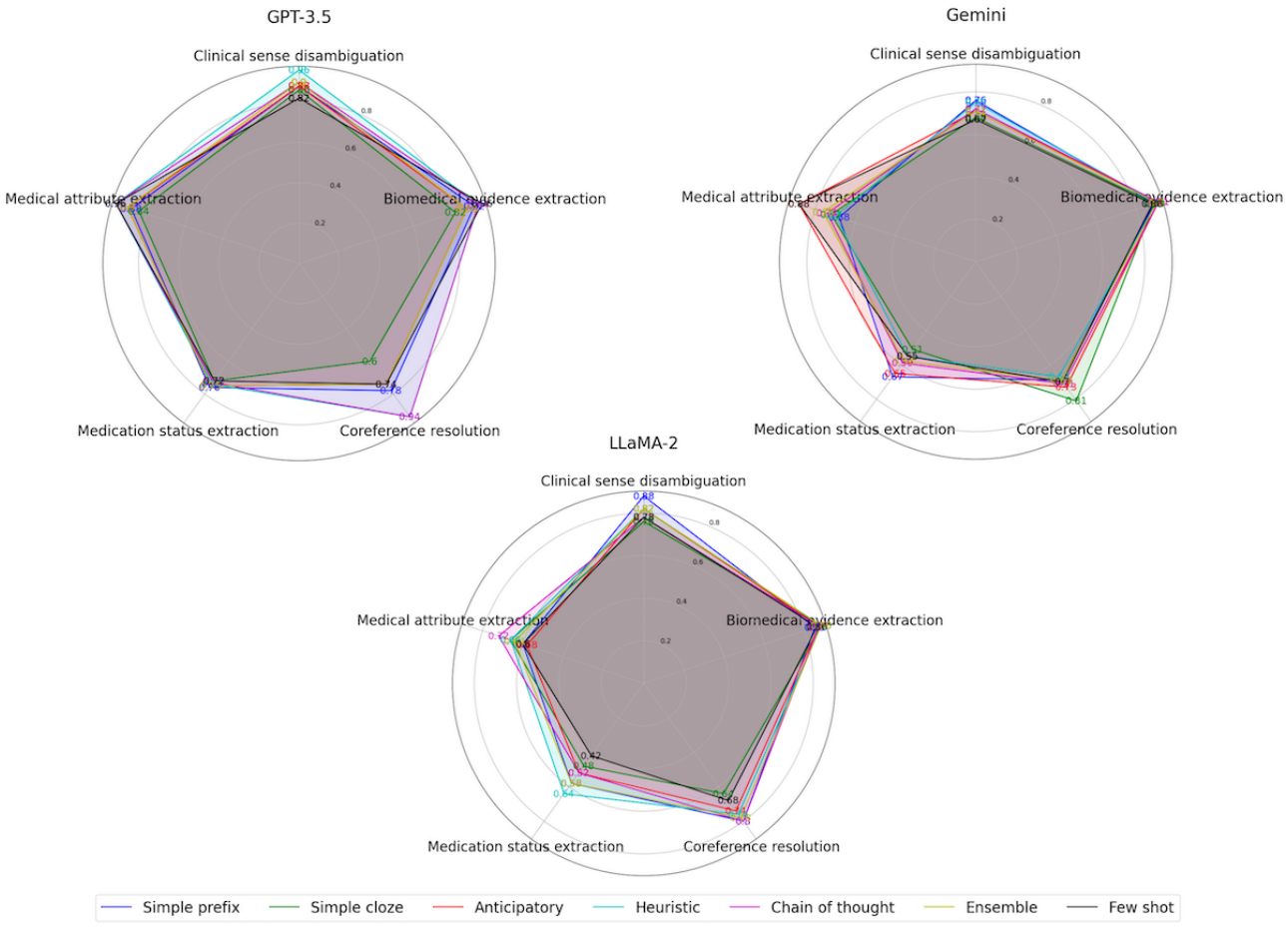
Prompt Optimization and Evaluation

For clinical sense disambiguation, the heuristic and prefix prompts consistently achieved the highest performance across all LLMs, significantly outperforming baselines such as BERT [30] and ELMO, with GPT-3.5 achieving an accuracy of 0.96, showcasing its advanced understanding of clinical context using appropriate prompting strategies. For biomedical evidence extraction, the heuristic and chain of thought prompts excelled across all LLMs in zero-shot setting. This indicates that these prompt types were able to provide enough information and constraints for the model to extract the evidence from the clinical note. GPT-3.5 achieved an accuracy of 0.94 with these prompt types, which was higher than any other model or prompt type combination. For coreference resolution, the chain of thought prompt type performed best among all prompt types with 2 LLMs—GPT-3.5 and LLaMA-2. This indicates that this prompt type was able to provide enough structure and logic for the model to resolve the coreference in the clinical note. GPT-3.5 displayed high accuracy with this prompt type, achieving an

accuracy of 0.94. For medication status extraction, simple prefix and heuristic prompts yielded good results across all LLMs. These prompt types were able to provide enough introduction or rules for the model to extract the status of the medication in relation to the patient or condition. GPT-3.5 excelled with these prompt types, achieving an accuracy of 0.76 and 0.74, respectively. For medication attribute extraction, we found that the chain of thought and heuristic prompts were effective across all LLMs. These prompt types were able to provide enough reasoning or rules for the model to extract and label the attributes of medications from clinical notes. Anticipatory prompts, however, had the best accuracy for Gemini among all the prompts. GPT-3.5 achieved an accuracy of 0.96 with these prompt types, which was higher than any other model or prompt type combination.

Thus, we can see that task-specific prompt tailoring is crucial for achieving high accuracy. Different tasks require different levels of information and constraints to guide the LLM to produce the desired output. The experiments show that heuristic, prefix, and chain of thought prompts are generally very effective for guiding the LLM to produce clear and unambiguous outputs. As shown in Figure 3, it is clear that GPT-3.5 is a superior and versatile LLM that can handle various clinical NLP tasks in zero-shot settings, outperforming other models in most cases.

Figure 3. Graphical comparison of prompt types in the 5 clinical natural language processing tasks used in this study.



Overall, the prompt-based approach has demonstrated remarkable superiority over traditional baseline models across all the 5 tasks. For clinical sense disambiguation, GPT-3.5's heuristic prompts achieved a remarkable accuracy of 0.96, showcasing a notable improvement over baselines such as BERT (0.42) and ELMO (0.55). In biomedical evidence extraction, GPT-3.5 again set a high standard with an accuracy of 0.94 using heuristic prompts, far surpassing the baseline performance of PubMedBERT-CRF at 0.35. Coreference resolution saw GPT-3.5 reaching an accuracy of 0.94 with chain of thought prompts, eclipsing the performance of existing methods such as Toshniwal et al [34] (0.69). In medication status extraction, GPT-3.5 outperformed the baseline ScispaCy (0.52) with an accuracy of 0.76 using simple prefix prompts. Finally, for medication attribute extraction, GPT-3.5's heuristic prompts achieved an impressive accuracy of 0.96, significantly higher than the ScispaCy baseline (0.70). These figures not only showcase the potential of LLMs in clinical settings but also set a foundation for future research to build upon, exploring even more sophisticated prompt engineering strategies and their implications for health care informatics.

Zero-Shot Versus Few-Shot Prompting

The performance of zero-shot prompting and few-shot prompting strategies was compared for each clinical NLP task. The same prompt types and LLMs were used as in the previous experiments, but some context was added to the input in the form of examples or explanations. Two examples or

explanations were used for each task (2-shot) depending on the complexity and variability of the task. Table 3 shows that few-shot prompting consistently improved the accuracy of all combinations for all tasks except for clinical sense disambiguation and medication attribute extraction, where some zero-shot prompt types performed better. We also observed some general trends across the tasks and models.

We found that few-shot prompting enhanced accuracy by providing limited context that aided complex scenario understanding. The improvement was more pronounced compared to simple cloze prompts, which had lower accuracy in most of the tasks. We also found that some zero-shot prompt types were very effective for certain tasks, even outperforming few-shot prompting. These prompt types used a rule-based or reasoning approach to generate sentences that contained definitions or examples of the target words or concepts, which helped the LLM to understand and match the context. For example, heuristic prompts achieved higher accuracy than few-shot prompting for clinical sense disambiguation and medication attribute extraction, while chain of thought prompts achieved higher accuracy than few-shot prompting for coreference resolution and medication attribute extraction. Alternatively, the clinical evidence extraction task likely benefits from additional context provided by few-shot examples, which can guide the model more effectively than the broader inferences made in zero-shot scenarios. This suggests that tasks requiring deeper contextual understanding might be better suited to few-shot learning approaches.

From these results, we can infer that LLMs can be effectively used for clinical NLP in a no-data scenario, where we do not have many publicly available data sets, by using appropriate zero-shot prompt types that guide the LLM to produce clear and unambiguous outputs. However, few-shot prompting can also improve the performance of LLMs by providing some context that helps the LLM to handle complex scenarios.

Other Observations

Ensemble Approaches

We experimented with an ensemble approach by combining outputs from multiple prompts using majority voting. The ensemble approach was not the best-performing strategy for any of the tasks, but it was better than the low-performing prompts. The ensemble approach was able to benefit from the diversity and complementarity of different prompt types and avoid some of the pitfalls of individual prompts. For example, for clinical sense disambiguation, the ensemble approach achieved an accuracy of 0.9 with GPT-3.5, which was the second best–performing prompt type. Similarly, for medication attribute extraction, the ensemble approach achieved an accuracy of 0.9 with GPT-3.5 and 0.76 with Gemini, which were close to the best single prompt type (anticipatory). However, the ensemble approach also had some drawbacks, such as inconsistency and noise. For tasks that required more specific or consistent outputs, such as coreference resolution, the ensemble approach did not improve the accuracy over the best single prompt type and sometimes even decreased it. This suggests that the ensemble approach may introduce ambiguity for tasks that require more precise or coherent outputs.

While the ensemble approach aims to reduce the variance introduced by individual prompt idiosyncrasies, our specific implementation observed instances where the combination of diverse prompt types introduced additional complexity. This complexity occasionally manifested as inconsistency and noise in the outputs contrary to our objective of achieving higher performance. Future iterations of this approach may include refinement of the prompt selection process to enhance consistency and further reduce noise in the aggregated outputs.

Impact of LLMs

Variations in performance were observed among different LLMs (Table 3). We found that GPT-3.5 generally outperformed Gemini and LLaMA-2 on most tasks. This suggests that GPT-3.5 has a better generalization ability and can handle a variety of clinical NLP tasks with different prompt types. However, Gemini and LLaMA-2 also showed some advantages over GPT-3.5 on certain tasks and prompt types. For example, Gemini achieved the highest accuracy of 0.81 with simple cloze prompts and LLaMA-2 achieved the highest accuracy of 0.8 with simple prefix prompts for coreference resolution. This indicates that Gemini and LLaMA-2 may have some domain-specific knowledge that can benefit certain clinical NLP tasks for specific prompt types.

Persona Patterns

Persona patterns are a way of asking the LLM to act like a persona or a system that is relevant to the task or domain. For example, one can ask the LLM to "act as a clinical NLP expert." This can help the LLM to generate outputs that are more appropriate and consistent with the persona or system. For example, one can use the following prompt for clinical sense disambiguation:

> *Act as a clinical NLP expert. Disambiguate the word "cold" in the following sentence: "She had a cold for three days."*

We experimented with persona patterns for different tasks and LLMs and found that they can improve the accuracy and quality of the outputs. Persona patterns can help the LLM to focus on the relevant information and constraints for the task and avoid generating outputs that are irrelevant or contradictory to the persona or system.

Randomness in Output

Most LLMs do not produce the output in the same format every time. There is inherent randomness in the outputs the LLMs produce. Hence, the prompts need to be specific in the way they are done for the task. Prompts are powerful when they are specific and if we use them in the right way.

Randomness in output can be beneficial or detrimental for different tasks and scenarios. In the clinical domain, randomness can introduce noise and errors in the outputs, which can make them less accurate and reliable for the users. For example, for tasks that involve extracting factual information, such as biomedical evidence extraction and medication status extraction, randomness can cause the LM to produce outputs that are inconsistent or contradictory with the input or context.

Guidelines and Suggestions for Optimal Prompt Selection

In recognizing the evolving nature of clinical NLP, we expand our discussion to contemplate the adaptability of our recommended prompt types and LM combinations across a wider spectrum of clinical tasks and narratives. This speculative analysis aims to bridge the gap between our current findings and their applicability to unexplored clinical NLP challenges, setting a foundation for future research to validate and refine these recommendations. In this section, we synthesize the main findings from our experiments and offer some practical advice for prompt engineering for zero-shot and few-shot clinical IE. We propose the following steps for selecting optimal prompts for different tasks and scenarios:

The first step is to identify the type of clinical NLP task, which can be broadly categorized into three types: (1) classification, (2) extraction, and (3) resolution. Classification tasks involve assigning a label or category to a word, phrase, or sentence in a clinical note, such as clinical sense disambiguation or medication status extraction. Extraction tasks involve identifying and extracting relevant information from a clinical note, such as biomedical evidence extraction or medication attribute extraction. Resolution tasks involve linking or matching entities or concepts in a clinical note, such as coreference resolution.

The second step is to choose the prompt type that is most suitable for the task type. We found that different prompt types have different strengths and weaknesses for different task types, depending on the level of information and constraints they provide to the LLM. Table 4 summarizes our findings and recommendations for optimal prompt selection for each task type.

The third step is to choose the LLM that is most compatible with the chosen prompt type. We found that different LLMs have different capabilities and limitations for different prompt types, depending on their generalization ability and domain-specific knowledge. Table 5 summarizes our findings and recommendations for optimal LLM selection for each prompt type.

The fourth step is to evaluate the performance of the chosen prompt type and LLM combination on the clinical NLP task using appropriate metrics such as accuracy, precision, recall, or F_1-score. If the performance is satisfactory, then the prompt engineering process is complete. If not, then the process can be repeated by choosing a different prompt type or LLM or by modifying the existing prompt to improve its effectiveness.

Table 4. Optimal prompt types for different clinical natural language processing task types.

Task type	Prompt type
Classification	Heuristic or prefix
Extraction	Heuristic or chain of thought
Resolution	Chain of thought

Table 5. Optimal language models for different prompt types.

Prompt type	Language model
Heuristic	GPT-3.5
Prefix	GPT-3.5 or LLaMA-2
Cloze	Gemini or LLaMA-2
Chain of thought	GPT-3.5
Anticipatory	Gemini

Discussion

Principal Findings

In this paper, we have presented a novel approach to zero-shot and few-shot clinical IE using prompt engineering. Various prompt types were evaluated across 5 clinical NLP tasks: clinical sense disambiguation, biomedical evidence extraction, coreference resolution, medication status extraction, and medication attribute extraction. The performance of different LLMs, GPT-3.5, Gemini, and LLaMA-2, was also compared. Our main findings are as follows:

1. Task-specific prompt tailoring is crucial for achieving high accuracy. Different tasks require different levels of information and constraints to guide the LLM to produce the desired output. Therefore, it is important to design prompts that are relevant and specific to the task at hand and avoid using generic or vague prompts that may confuse the model or lead to erroneous outputs.
2. Heuristic prompts are generally very effective for guiding the LLM to produce clear and unambiguous outputs. These prompts use a rule-based approach to generate sentences that contain definitions or examples of the target words or concepts, which help the model to understand and match the context. Heuristic prompts are especially useful for tasks that involve disambiguation, extraction, or classification of entities or relations.
3. Chain of thought prompts are also effective for guiding the LLM to produce logical and coherent outputs. These prompts use a multistep approach to generate sentences that contain a series of questions and answers that resolve the

task in the context. Chain of thought prompts are especially useful for tasks that involve reasoning, inference, or coreference resolution.

4. Few-shot prompting can improve the performance of LLMs by providing some context that helps the model to handle complex scenarios. Few-shot prompting can be done by adding some examples or explanations to the input depending on the complexity and variability of the task. Few-shot prompting can enhance accuracy by providing limited context that aids complex scenario understanding. The improvement is more pronounced compared to simple prefix and cloze prompts, which had lower accuracy in most of the tasks.
5. Ensemble approaches can also improve the performance of LLMs by combining outputs from multiple prompts using majority voting. Ensemble approaches can leverage the strengths of each prompt type and reduce the errors of individual prompts. Ensemble approaches are especially effective for tasks that require multiple types of information or reasoning, such as biomedical evidence extraction and medication attribute extraction.

It is noteworthy that context size has a significant impact on the performance of LLMs in zero-shot IE [36]. In the scope of this study, we have avoided the context size dependence on performance, as it is a complex issue that requires careful consideration.

This study serves as an initial exploration into the efficacy of prompt engineering in clinical NLP, providing foundational insights rather than exhaustive guidelines. Given the rapid advancements in generative artificial intelligence and the complexity of clinical narratives, we advocate for continuous empirical testing of these prompt strategies across diverse clinical tasks and data sets. This approach will not only validate the generalizability of our findings but also uncover new avenues for enhancing the accuracy and applicability of LLMs in clinical settings.

Limitations

In this study, we primarily focused on exploring the capabilities and versatility of generative LLMs in the context of zero-shot and few-shot learning for clinical NLP tasks. Our approach also has some limitations that we acknowledge in this work. First, it relies on the quality and availability of pretrained LLMs, which may vary depending on the domain and task. As LLMs are rapidly evolving, some parts of the prompt engineering discipline may be timeless, while some parts may evolve and adapt over time as different capabilities of models evolve. Second, it requires a lot of experimentation and iteration to optimize prompts for different applications, which may be iterative and time-consuming. However, once optimal prompts are identified, the approach offers time savings in subsequent applications by reusing these prompts or making minor adjustments for similar tasks. We may not have explored all the possible combinations and variations of prompts that could potentially improve the performance of the clinical NLP tasks. Third, the LLMs do not release the details of the data set that they were trained on. Hence, the high accuracy could be because the models would have already seen the data during training and not because of the effectiveness of the prompts.

Future Work

We plan to address these challenges and limitations in our future work. We aim to develop more systematic and automated methods for prompt design and evaluation, such as using prompt-tuning or meta-learning techniques. We also aim to incorporate more domain knowledge or external resources into the prompts or the LLMs, such as using ontologies, knowledge graphs, or databases. We also aim to incorporate more quality control or error correction mechanisms into the prompts or the LLMs, such as using adversarial examples, confidence scores, or human feedback.

Conclusions

In this paper, we have benchmarked different prompt engineering techniques for both zero-shot and few-shot clinical NLP tasks. Two new types of prompts, heuristic and ensemble prompts, were also conceptualized and proposed. We have demonstrated that prompt engineering can enable the use of pretrained LMs for various clinical NLP tasks without requiring any fine-tuning or additional data. We have shown that task-specific prompt tailoring, heuristic prompts, chain of thought prompts, few-shot prompting, and ensemble approaches can improve the accuracy and quality of the outputs. We have also shown that GPT-3.5 is very adaptable and precise across all tasks and prompt types, while Gemini and LLaMA-2 may have some domain-specific advantages for certain tasks and prompt types.

We believe that a prompt-based approach has several benefits over existing methods for clinical IE. It reduces the cost and time in the initial phases of clinical NLP application development, where prompt-based methods offer a streamlined alternative to the conventional data preparation and model training processes. It is flexible and adaptable, as it can be applied to various clinical NLP tasks with different prompt types and LLMs. It is interpretable and explainable, as it uses natural language prompts that can be easily understood and modified by humans.

Acknowledgments

This work was supported by the National Institutes of Health (awards U24 TR004111 and R01 LM014306). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

SS conceptualized, designed, and organized this study; analyzed the results; and wrote, reviewed, and revised the paper. MK and AS-M analyzed the results, and wrote, reviewed, and revised the paper. SV wrote, reviewed, and revised the paper. YW conceptualized, designed, and directed this study and wrote, reviewed, and revised the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Prompts for clinical natural language processing tasks.
[DOCX File , 31 KB - medinform_v12i1e55318_app1.docx]

References

1. 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]
2. Landolsi MY, Hlaoua L, Romdhane LB. Information extraction from electronic medical documents: state of the art and future research directions. Knowl Inf Syst 2023;65(2):463-516 [FREE Full text] [doi: 10.1007/s10115-022-01779-1] [Medline: 36405956]
3. Sivarajkumar S, Wang Y. HealthPrompt: a zero-shot learning paradigm for clinical natural language processing. AMIA Annu Symp Proc 2022;2022:972-981 [FREE Full text] [Medline: 37128372]
4. Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, et al. Rethinking the role of demonstrations: what makes in-context learning work? 2022 Presented at: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; December 7-11, 2022; Abu Dhabi, United Arab Emirates p. 11048-11064 URL: https://aclanthology.org/2022.emnlp-main.759/ [doi: 10.18653/v1/2022.emnlp-main.759]
5. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with chatGPT. ArXiv Preprint posted online on February 21, 2023 [FREE Full text] [doi: 10.48550/arXiv.2302.11382]
6. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Red Hook, NY: Curran Associates, Inc; 2022:27730-27744.
7. Gemini Team Google. Gemini: a family of highly capable multimodal models. ArXiv Preprint posted online on December 19, 2023 [FREE Full text] [doi: 10.48550/arXiv.2312.11805]
8. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. ArXiv Preprint posted online on July 28, 2023 [FREE Full text] [doi: 10.48550/arXiv.2307.09288]
9. Chen Q, Du J, Hu Y, Keloth VK, Peng X, Raja K, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. ArXiv Preprint posted online on May 10, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.16326]
10. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt engineering for healthcare: methodologies and applications. ArXiv Preprint posted online on April 28, 2023 [FREE Full text] [doi: 10.48550/arXiv.2304.14670]
11. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. ArXiv Preprint posted online on March 29, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.16416]
12. Yuan C, Xie Q, Ananiadou S. Zero-shot temporal relation extraction with chatGPT. 2023 Presented at: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; July 13, 2023; Toronto, Canada p. 92-102 URL: https://aclanthology.org/2023.bionlp-1.7/ [doi: 10.18653/v1/2023.bionlp-1.7]
13. Li X, Liang L. Prefix-tuning: optimizing continuous prompts for generation. 2021 Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); August 1-6, 2021; Virtual Event p. 4582-4597 URL: https://aclanthology.org/2021.acl-long.353/ [doi: 10.18653/v1/2021.acl-long.353]
14. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023;55(9):1-35 [FREE Full text] [doi: 10.1145/3560815]
15. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems 35 (NeurIPS 2022). Red Hook, NY: Curran Associates Inc; 2022:24824-24837.
16. Hancock B, Bordes A, Mazare PE, Weston J. Learning from dialogue after deployment: feed yourself, chatbot!. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy p. 3667-3684 URL: https://aclanthology.org/P19-1358/ [doi: 10.18653/v1/p19-1358]
17. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. J Biomed Inform 2009;42(5):923-936 [FREE Full text] [doi: 10.1016/j.jbi.2009.07.007] [Medline: 19646551]

18. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. : Association for Computational Linguistics; 2022 Presented at: The 2022 Conference on Empirical Methods in Natural Language Processing; December 7-11, 2022; Abu Dhabi, United Arab Emirates p. 1998-2022 URL: https://aclanthology.org/2022.emnlp-main.130.pdf [doi: 10.18653/v1/2022.emnlp-main.130]
19. Wu Y, Xu J, Zhang Y, Xu H. Clinical abbreviation disambiguation using neural word embeddings. 2015 Presented at: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015); July 30, 2015; Beijing, China p. 171-176 URL: https://aclanthology.org/W15-3822.pdf [doi: 10.18653/v1/w15-3822]
20. Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Iorio A, et al. Machine learning approaches to retrieve high-quality, clinically relevant evidence from the biomedical literature: systematic review. JMIR Med Inform 2021;9(9):e30401 [FREE Full text] [doi: 10.2196/30401] [Medline: 34499041]
21. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012;19(5):786-791 [FREE Full text] [doi: 10.1136/amiajnl-2011-000784] [Medline: 22366294]
22. Denny JC, Peterson JF, Choma NN, Xu H, Miller RA, Bastarache L, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010;17(4):383-388 [FREE Full text] [doi: 10.1136/jamia.2010.004804] [Medline: 20595304]
23. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Saf 2019;42(1):99-111 [FREE Full text] [doi: 10.1007/s40264-018-0762-z] [Medline: 30649735]
24. Chen Y, Wu X, Chen M, Song Q, Wei J, Li X, et al. Dynamic text categorization of search results for medical class recognition in real world evidence studies in the Chinese language. : Association for Computing Machinery, Presented at: Proceedings of the International Conference on Bioinformatics and Computational Intelligence (ICBCI 2017); 2017; Beijing, China p. 40-48. [doi: 10.1145/3135954.3135962]
25. Mallick PK, Balas VE, Bhoi AK, Zobaa AF. Cognitive Informatics and Soft Computing Proceeding of CISC 2017, Advances in Intelligent Systems and Computing (AISC, Volume 768). New York: Springer Verlag; 2019.
26. Ananiadou S, Lee D, Xu H, Song M. DTMBIO'12—The Proceedings of the Sixth ACM International Workshop on Data and Text Mining in Biomedical Informatics. 2012 Presented at: 6th ACM International Workshop on Data and Text Mining in Biomedical Informatics, DTMBIO 2012, in Conjunction with the 21st ACM International Conference on Information and Knowledge Management, CIKM 2012; 2012; New York URL: https://dl.acm.org/action/showFmPdf?doi=10.1145%2F2390068 [doi: 10.1145/2396761.2398758]
27. Elghandour I, State R, Brorsson M, Le L, Antonopoulos N, Xie Y, et al. IEEE/ACM International Symposium on Big Data Computing (BDC). 2016 Presented at: 2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT); December 6-9, 2016; Shanghai, China URL: https://ieeexplore.ieee.org/xpl/conhome/7876287/proceeding [doi: 10.1109/bdcat.2018.00008]
28. Moon S, Pakhomov S, Liu N, Ryan JO, Melton GB. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. J Am Med Inform Assoc 2014;21(2):299-307 [FREE Full text] [doi: 10.1136/amiajnl-2012-001506] [Medline: 23813539]
29. Nye B, Li JJ, Patel R, Yang Y, Marshall I, Nenkova A, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. 2018 Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 15-20, 2018; Melbourne, Australia p. 197-207 URL: https://aclanthology.org/P18-1019/ [doi: 10.18653/v1/p18-1019]
30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint posted online on May 24, 2019 [FREE Full text]
31. Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I, Jarkiewicz M, et al. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Res 2021;304:114135. [doi: 10.1016/j.psychres.2021.114135] [Medline: 34343877]
32. 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 Healthcare 2021;3(1):1-23. [doi: 10.1145/3458754]
33. Adams G, Ketenci M, Bhave S, Perotte A, Elhadad N. Zero-shot clinical acronym expansion via latent meaning cells. Proc Mach Learn Res 2020;136:12-40 [FREE Full text] [Medline: 34790898]
34. Toshniwal S, Xia P, Wiseman S, Livescu K, Gimpel K. On generalization in coreference resolution. ArXiv Preprint posted online on September 20, 2021 [FREE Full text] [doi: 10.18653/v1/2021.crac-1.12]
35. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: fast and robust models for biomedical natural language processing. ArXiv Preprint posted online on October 9, 2019 [FREE Full text] [doi: 10.18653/v1/w19-5034]
36. Sivarajkumar S, Wang Y. Evaluation of healthprompt for zero-shot clinical text classification. 2023 Presented at: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI); June 26-29, 2023; Houston, TX, USA. [doi: 10.1109/ichi57859.2023.00081]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
ELMO: Embeddings From Language Models
IE: information extraction
LLM: large language model
LM: language model
NLP: natural language processing
PubMedBERT-CRF: PubMedBERT-Conditional Random Field

Edited by C Lovis; submitted 08.12.23; peer-reviewed by J Zaghir, M Torii, J Zheng; comments to author 04.02.24; revised version received 20.02.24; accepted 24.02.24; published 08.04.24.

Please cite as:
Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y
An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study
JMIR Med Inform 2024;12:e55318
URL: https://medinform.jmir.org/2024/1/e55318
doi:10.2196/55318
PMID:38587879

Original Paper

Evaluating ChatGPT-4's Diagnostic Accuracy: Impact of Visual Data Integration

Takanobu Hirosawa[1*], MD, PhD; Yukinori Harada[1*], MD, PhD; Kazuki Tokumasu[2*], MD, PhD; Takahiro Ito[3*], MD; Tomoharu Suzuki[4*], MD; Taro Shimizu[1*], MD, MSc, MPH, MBA, PhD

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Shimotsuga, Japan

[2]Department of General Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

[3]Satsuki Home Clinic, Tochigi, Japan

[4]Department of Hospital Medicine, Urasoe General Hospital, Okinawa, Japan

[*]all authors contributed equally

Corresponding Author:
Takanobu Hirosawa, MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
880 Kitakobayashi, Mibu-cho
Shimotsuga, 321-0293
Japan
Phone: 81 282 87 2498
Email: hirosawa@dokkyomed.ac.jp

Abstract

Background: In the evolving field of health care, multimodal generative artificial intelligence (AI) systems, such as ChatGPT-4 with vision (ChatGPT-4V), represent a significant advancement, as they integrate visual data with text data. This integration has the potential to revolutionize clinical diagnostics by offering more comprehensive analysis capabilities. However, the impact on diagnostic accuracy of using image data to augment ChatGPT-4 remains unclear.

Objective: This study aims to assess the impact of adding image data on ChatGPT-4's diagnostic accuracy and provide insights into how image data integration can enhance the accuracy of multimodal AI in medical diagnostics. Specifically, this study endeavored to compare the diagnostic accuracy between ChatGPT-4V, which processed both text and image data, and its counterpart, ChatGPT-4, which only uses text data.

Methods: We identified a total of 557 case reports published in the *American Journal of Case Reports* from January 2022 to March 2023. After excluding cases that were nondiagnostic, pediatric, and lacking image data, we included 363 case descriptions with their final diagnoses and associated images. We compared the diagnostic accuracy of ChatGPT-4V and ChatGPT-4 without vision based on their ability to include the final diagnoses within differential diagnosis lists. Two independent physicians evaluated their accuracy, with a third resolving any discrepancies, ensuring a rigorous and objective analysis.

Results: The integration of image data into ChatGPT-4V did not significantly enhance diagnostic accuracy, showing that final diagnoses were included in the top 10 differential diagnosis lists at a rate of 85.1% (n=309), comparable to the rate of 87.9% (n=319) for the text-only version (P=.33). Notably, ChatGPT-4V's performance in correctly identifying the top diagnosis was inferior, at 44.4% (n=161), compared with 55.9% (n=203) for the text-only version (P=.002, χ^2 test). Additionally, ChatGPT-4's self-reports showed that image data accounted for 30% of the weight in developing the differential diagnosis lists in more than half of cases.

Conclusions: Our findings reveal that currently, ChatGPT-4V predominantly relies on textual data, limiting its ability to fully use the diagnostic potential of visual information. This study underscores the need for further development of multimodal generative AI systems to effectively integrate and use clinical image data. Enhancing the diagnostic performance of such AI systems through improved multimodal data integration could significantly benefit patient care by providing more accurate and comprehensive diagnostic insights. Future research should focus on overcoming these limitations, paving the way for the practical application of advanced AI in medicine.

(JMIR Med Inform 2024;12:e55627) doi:10.2196/55627

KEYWORDS

artificial intelligence; large language model; LLM; LLMs; language model; language models; ChatGPT; GPT; ChatGPT-4V; ChatGPT-4 Vision; clinical decision support; natural language processing; decision support; NLP; diagnostic excellence; diagnosis; diagnoses; diagnose; diagnostic; diagnostics; image; images; imaging

Introduction

Diagnostic Excellence

Diagnostic excellence involves accurately and efficiently diagnosing a wide range of conditions [1]. Achieving this requires a multifaceted approach [2], including effective collaboration among medical professionals, patients, families, and clinical decision support systems (CDSSs). Each plays a pivotal role, as follows: medical professionals bring their expertise and judgment, patients and families provide essential health information and context, and CDSSs offer data-driven insights, enhancing the collective decision-making process.

CDSSs for Diagnostic Excellence

CDSSs are computer-based tools that assist medical professionals in a wide range of clinical decisions, including diagnosis, treatment planning, medication ordering, preventive care, and patient education [3]. Research has shown that CDSS interventions significantly improve diagnostic accuracy [4], a key aspect of diagnostic excellence [5]. For instance, interventions involving a CDSS in the diagnosis of common chronic diseases demonstrated significant improvements [6]. Accurate diagnosis entails more than identifying a disease; it involves understanding the patient's unique health context, ensuring timely and appropriate treatment, reducing misdiagnosis risk, and ultimately improving patient outcomes [7]. In the rapidly evolving health care environment, maintaining high standards of diagnostic precision becomes increasingly crucial.

Artificial Intelligence in Medicine

CDSSs are broadly categorized into 2 types [3]: knowledge-based systems, which are grounded in medical guidelines and expert knowledge; and non–knowledge-based systems, using artificial intelligence (AI) or statistical pattern recognition for clinical data analysis.

The integration of AI into clinical settings is advancing rapidly. AI systems in medicine range from assisting in diagnostic imaging and analysis to optimizing patient treatment plans [8,9]. These systems are being increasingly adopted in hospitals and clinics [10], significantly contributing to enhanced diagnostic accuracy and efficiency.

However, the integration of AI into clinical settings brings transformative potential but also faces several hurdles. Challenges include ensuring data privacy [11], addressing the lack of large and diverse training data sets, and maintaining the interpretability of AI-generated recommendations to align with ethical standards [12,13]. Real-world obstacles, such as resistance from health care professionals due to trust issues in AI's diagnostic suggestions, underscore the complexity of AI integration into clinical practice.

Advancements in Large Language Models

A notable advancement in AI is the use of large language models (LLMs). As a subset of non–knowledge-based systems, LLMs are specialized forms of generative AI systems that process and generate human-like text based on extensive textual data training [14]. They are adept at tasks like translation, summarization, and even creative writing. In clinical practice, generative AI systems using LLMs have shown promise in summarizing patient history, integrating medical records, analyzing complex data streams, and enhancing communication between patients and medical professionals [15,16], demonstrating their utility in handling complex medical language and concepts. Such advancements not only improve the efficiency of medical documentation but also offer novel approaches to generating differential diagnoses, showcasing the innovative application of LLMs in clinical settings.

Multimodal Artificial Intelligence in Diagnostics

Integrating multimodal data, including text and images, presents technical challenges. Successful integration in other fields, such as autonomous driving technologies that combine multisensory observation data to navigate [17], offers a potential model for health care. Recent developments in generative AI systems, including Google Gemini (previous Google Bard [18]) and ChatGPT-4 with vision (ChatGPT-4V), have enabled the processing of both text and image data. This integration is essential for providing a comprehensive clinical overview. Although effectively combining data from different data sources remains a challenge, the development of multimodal AI models that incorporate data across modalities enabled broad applications that include personalized medicine and digital health [19]. For example, a multimodal model developed from the combination of images and health records could classify pulmonary embolism [20]. Another multimodal model could differentiate between common respiratory failure [21]. Among publicly available generative AI systems, the ChatGPT series, particularly ChatGPT-4V, developed by OpenAI and released in September 2023, stands out [22,23]. It accepts both text and image data [24,25], demonstrating impressive performance in various applications.

Preliminary studies in various fields, including medicine [26-28] and others [29-31] have shown the effectiveness of ChatGPT-4V. Some of these studies have highlighted its efficacy in interpreting medical images [26,28], though they were limited in scope. However, clinical image data includes a wide range of elements, from physical examinations to various investigation results. The full impact of image data integration on diagnostic accuracy is yet to be thoroughly explored.

Study Objectives

This study directly addressed the gaps identified in the current understanding of multimodal AI's application in clinical diagnostics. By comparing the diagnostic accuracy of

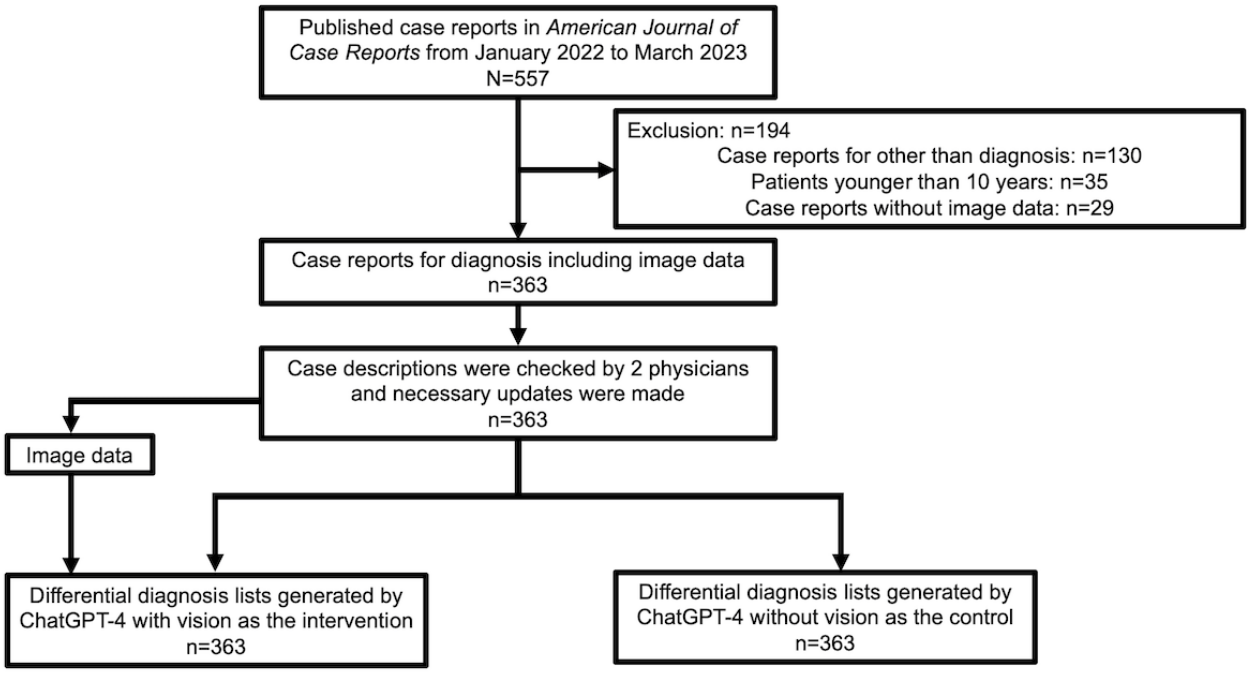
ChatGPT-4V and without vision across detailed case reports, and examining the impact of image data integration, we aimed to provide concrete evidence on the value and challenges of incorporating generative AI into clinical flows. Our objectives were shaped by the need to better understand how multimodal AI can be optimized to support diagnostic excellence, ultimately contributing to the advancement of medical diagnostics through technology.

Methods

Overview

We conducted an experimental study to assess the diagnostic accuracy of multimodal generative AI systems using data from a large number of case reports. This study was conducted in the Department of Diagnostic and Generalist Medicine (General Internal Medicine) at Dokkyo Medical University. This study involved several steps: preparing the data set and control, preparing image data, generating differential diagnosis lists by ChatGPT-4V, and evaluating the diagnostic accuracy of these differential diagnosis lists. A flow chart of the study's methodology is presented in Figure 1.

Figure 1. Study design.



Ethical Considerations

This study used published case reports, and thus ethical committee approval was not applicable.

Preparing Data Set and Control

We used the data set from our previous study (T Hirosawa, Y Harada, K Mizuta, T Sakamoto, K Tokomasu, T Shimizu, unpublished data, November 2023). The data set comprised case descriptions and final diagnoses, sourced from the *American Journal of Case Reports*, spanning January 2022 to March 2023. This peer-reviewed journal covers diagnostically challenging case reports from various medical fields. A total of 557 case reports were identified. The exclusion criteria were carefully chosen based on previous studies for CDSSs [32] and ChatGPT-4V [28] to ensure the focus remained on diagnostically challenging adult cases with relevant image data. Specifically, cases were excluded for the following reasons: nondiagnosis (130 cases), patients younger than 10 years (35 cases), and the absence of image data (29 cases). The included case reports were refined into case descriptions by the primary researcher (TH) and double-checked by another researcher (YH). From the included case reports, we extracted a case description until the final diagnosis was made in the "case report" section. We removed sentences that directly assessed the diagnosis to minimize bias in generating differential diagnoses. This step ensures that the differential diagnoses generated by ChatGPT-4 are based solely on the unbiased clinical presentation of the case. After brush-up, we formatted these case descriptions for input into ChatGPT-4. A typical case description included demographic information, chief complaints, history of present illness, results of physical examinations, and investigative findings leading to diagnoses. The final diagnoses were typically determined by the authors of the case reports. For example, in a case report titled "Levofloxacin-Associated Bullous Pemphigoid in a Hemodialysis Patient After Kidney Transplant Failure" [33] we extracted from "A 27-year-old female with hemodialysis was admitted for evaluation of a worsening bullous rash and shortness of breath over the last 3 days..." to "...Although the swab PCR test for VZV and HSV was negative, there was still concern about disseminated herpes zoster, as the patient was immunosuppressed" as a case description.

Additionally, the final diagnosis was levofloxacin-associated bullous pemphigoid.

In the next step, we used ChatGPT-4 without vision to develop the top 10 differential diagnosis lists based on the data of case descriptions. Two expert physicians independently evaluated whether the final diagnosis was included in the lists, and any discrepancies were resolved through discussion. Therefore, the differential diagnosis lists and data of physicians' evaluation of the lists from a total of 363 case reports were included as the control in this study.

Preparing Image Data

All figures and tables of included case descriptions were standardized to a resolution of 96 dots per inch in JPEG format to balance detail with file size, facilitating efficient processing by ChatGPT-4V without compromising the quality necessary for accurate diagnostic inference. When multiple figures or tables were present in a case description, they were compiled into a single JPEG file, each annotated with a file number in the upper-left corner. If image data exceeded the upload size limit, the images were resized to half their original size while preserving image quality, using the Preview application (version 11.0; Apple Inc) on a Mac computer.

Generating Differential Diagnosis Lists by ChatGPT-4V

We used ChatGPT-4V, a multimodal generative AI system developed by OpenAI, from October 30, 2023, to November 9, 2023. Additional training or reinforcement for diagnosis was not performed. The prompt was constructed as follows: "Identify the top 10 suspected illnesses based on the attached files with file names indicated in the left upper corner of each image, and the provided case description. List these illnesses using only their names, without providing any reasoning AND describe the proportion of the case description and the provided files to develop your suspected illness list (case description + all files = 100%): (copy and paste the case descriptions)." This design was intended to explicitly guide ChatGPT-4V to not only generate a list of possible diagnoses but also reflect on how each type of data influenced its conclusions, providing insights into the AI's diagnostics process. Apart from the prompt and file names, the text data input to ChatGPT-4V remained the same as the control, ChatGPT-4 without vision. The first generated list was used as the differential diagnosis list. The chat history was cleared before entering each new case description. Moreover, the data control settings for chat history were disabled. The details of ChatGPT-4V and ChatGPT-4 without vision are shown in Table 1.

Table 1. The details of ChatGPT-4 with vision and ChatGPT-4 without vision in this study.

Details	ChatGPT-4 with vision (intervention) [24]	ChatGPT-4 without vision (control) [22]
Short name	ChatGPT-4V	ChatGPT-4
Prompt	Identify the top 10 suspected illnesses based on the attached files with file names indicated in the left upper corner of each image, and the provided case description. List these illnesses using only their names, without providing any reasoning AND describe the proportion of the case description and the provided files to develop your suspected illness list (case description + all files =100%): (copy and paste the case descriptions)	Tell me the top 10 suspected illnesses for the following case: (copy and paste the case descriptions)
Text input	Same case descriptions with the above prompt and referred file number	Same case descriptions with the above prompt
Image input	Image data in JPEG format with a resolution of 96 dots per inch	No image data
Output	The top 10 differential diagnosis lists and the proportion of weight between text data and image data contributing to development of the differential diagnosis list	The top 10 differential diagnosis lists
Evaluations	By 2 independent physicians; any discrepancies were resolved by another physician	By 2 independent physicians; any discrepancies were resolved by another physician
Release date	September 2023	March 2023
Access date	From October 30, 2023, to November 9, 2023	From June 22, 2023, to June 29, 2023
Data control for chat history	Off	Off

Evaluation for Differential Diagnosis Lists by Physicians

Two expert physicians, TI and T Suzuki, independently evaluated whether the final diagnoses were included in the differential diagnosis lists. The evaluation was binary, with 1 indicating inclusion and 0 indicating exclusion. A score of 1 indicated that the differential closely matched the final diagnoses. This close match was defined not merely by the presence of the correct diagnosis within the list but by the relevance and clinical appropriateness of the differentials in relation to the final diagnosis. A score of 1 indicated that AI-generated differentials were clinically relevant and could potentially lead to appropriate interventions, thereby aligning with patient safety and standards [34]. Additionally, evaluators ranked the match of differential to the final diagnoses.

Conversely, a score of 0 was given if the differential diagnosis list significantly differed from the final diagnosis, indicating a lack of clinical relevance or potential misdirection in a real-world diagnostic scenario. Any discrepancies were resolved by another expert physician (KT), ensuring objective and consistent evaluation across all included case reports.

Outcome

The study assessed the diagnostic accuracy of ChatGPT-4V, as an intervention and compared it to ChatGPT-4 without vision as a control. The primary outcome was defined as the ratio of cases where the final diagnoses were included within the top 10 differential diagnosis lists. The secondary outcome is defined as the ratio of cases where the final diagnoses were included as top diagnosis. These outcomes were chosen to quantitatively measure diagnostic accuracy and the effectiveness of image data integration in enhancing ChatGPT-4's diagnostics.

Additionally, we assessed the contributing weight between text data (case descriptions) and image data (files) in developing the differential diagnosis lists, as reported by ChatGPT-4V. The total contribution from both elements was set to 100%. Specifically, we analyzed how much the text and image data individually contributed to the formulation of the differential diagnosis list. For example, if the text data (case description) contributed 60% and the image data contributed 40%, the total would sum up to 100%. This method allowed for a comprehensive understanding of the relative impact of textual and image data on AI diagnostics.

Statistical Analysis

For analysis, R (version 4.2.2; R Foundation for Statistical Computing) was used. We present continuous variables as medians and IQRs to accurately reflect the distribution of data. We presented categorical or binary variables as numbers and percentages. Additionally, we used χ^2 tests to compare categorical variables, setting the significance level at a *P* value <.05. The choice of χ^2 tests for comparing categorical variables was based on their ability to handle binary and categorical data effectively, providing a robust measure of association between diagnostic outcomes and ChatGPT-4 with or without vision.

To quantify the impact of each factor on the likelihood of accurate diagnosis inclusion, an univariable logistic regression model was applied. This model allows for the exploration of potential predictors of diagnostic accuracy, offering insights into how different data types contribute to ChatGPT-4's diagnostic processes. For the logistic regression model, the primary and secondary outcomes were treated as binary dependent variables: presence (1) or absence (0) of the correct diagnosis within the top 10 differential diagnosis lists and as the top diagnosis, respectively. Independent variables included the proportion of image data weight, the presence (1) or absence (0) of specific types of image data (eg, computed tomography [CT] images, pathological specimens, laboratory data, magnetic resonance imaging [MRI] scans, and X-ray images), and the number of characters in the text data. Odds ratios (ORs) and associated 95% CIs were used to estimate the relative risks of potential predictors of the final diagnosis included within the top 10 differential diagnosis lists in the univariable logistic regression model.

Results

Case Descriptions and Image Data Profile

A total of 363 case descriptions with additional image data, such as figures or tables, were included. ChatGPT-4V generated the differential diagnosis lists for all case descriptions. Representative final diagnosis, image data, and differential diagnosis lists generated by ChatGPT-4V and ChatGPT-4 without vision are shown in Table 2. The cases included in this study, along with the differential diagnosis lists generated by ChatGPT-4V and without vision, are shown in Multimedia Appendix 1.

Table 2. Representative final diagnoses, image data, and differential diagnosis lists generated by ChatGPT-4 with vision and ChatGPT-4 without vision.

Final diagnoses	ChatGPT-4 with vision (intervention)			ChatGPT-4 without vision (control)
	Image data	Image data weight[a], %	Differential diagnosis lists	Differential diagnosis lists
Olanzapine intoxication	Laboratory data	20%	1. Hospital-acquired pneumonia 2. Sepsis 3. Acute kidney injury 4. Respiratory failure 5. Intensive care unit delirium 6. Hypotension 7. Opioid intoxication 8. Benzodiazepine intoxication 9. Hyperglycemia 10. Cardiovascular complications after surgery	1. Zolpidem overdose or prolonged effect 2. Benzodiazepine overdose 3. Olanzapine overdose or side effect 4. Interaction between zolpidem and olanzapine 5. Delirium 6. Other drug overdose 7. Severe sepsis 8. Neuroleptic malignant syndrome 9. Acute cholinergic crisis 10. Acute stroke
Secondary syphilis	Laboratory data and X-ray images	40%	1. Osteoarthritis 2. Rheumatoid arthritis 3. Gout 4. Pseudogout 5. Systemic lupus erythematosus 6. Psoriatic arthritis 7. Reactive arthritis 8. Osteoporosis 9. Septic arthritis 10. Ankylosing spondylitis	1. Secondary syphilis 2. Reactive arthritis 3. Rheumatoid arthritis 4. Osteoarthritis 5. Gout/pseudogout 6. Psoriatic arthritis 7. Lyme disease 8. Systemic lupus erythematosus 9. Infectious arthritis 10. Vasculitis (eg, granulomatosis with polyangiitis and microscopic polyangiitis)
Sarcomatoid carcinoma	Computed tomography image and pathological specimens	30%	1. Colorectal cancer 2. Adenocarcinoma of the colon 3. Malignant mesothelioma 4. Gastrointestinal stromal tumor 5. Carcinoid tumor of the appendix 6. Lymphoma 7. Desmoid tumor 8. Metastatic carcinoma 9. Leiomyosarcoma 10. Neuroendocrine tumor of the colon	1. Colorectal carcinoma 2. Metastatic malignancy to the colon 3. Peritoneal mesothelioma 4. Sarcomatoid carcinoma 5. Malignant mixed Mullerian tumor (carcinosarcoma) 6. Gastrointestinal stromal tumor 7. Leiomyosarcoma 8. Colonic lymphoma 9. Malignant peripheral nerve sheath tumors 10. Undifferentiated/unclassified malignancies

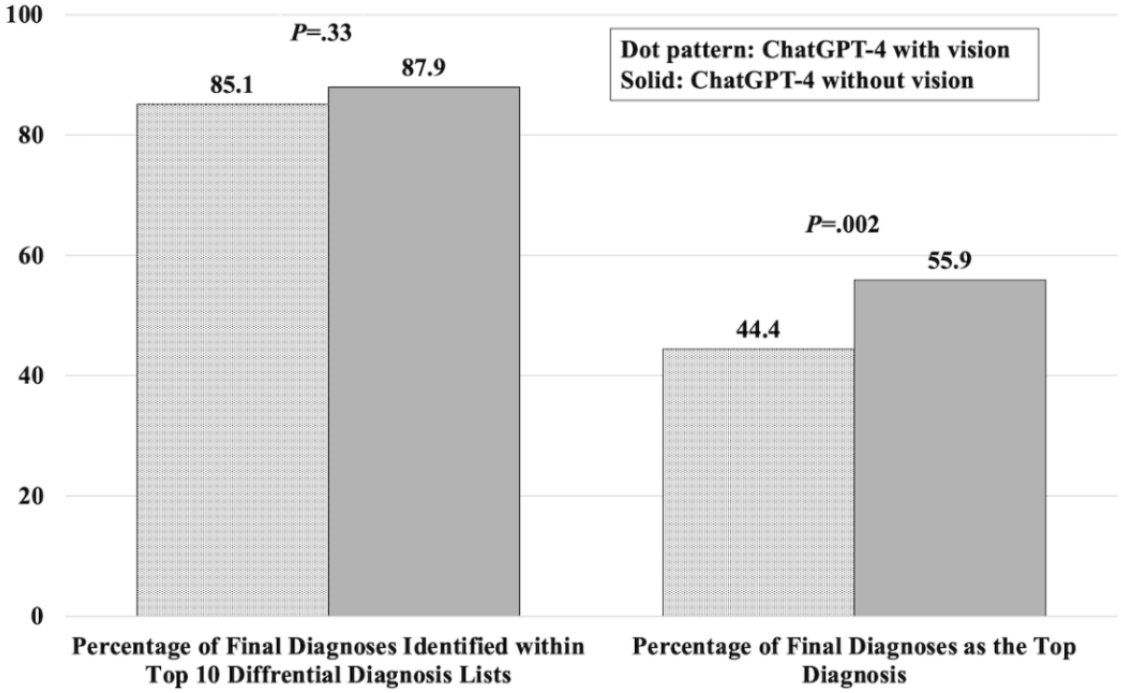
[a]The proportion of image data weight contributing to development of the differential-diagnosis lists.

Among these, the 25th percentile, median, and 75th percentile number of characters in the text data were 1971, 2683, and 3442, respectively. The maximum and minimum number of characters in text data were 7148 and 465, respectively. Regarding image data, CT images, pathological specimens, laboratory data, MRI scans, and X-ray images were included in 163, 124, 98, 77, and 70 case descriptions, respectively. The details of image data are shown in Multimedia Appendix 2.

Diagnostic Performance

For the primary outcome, the rate of final diagnoses within the top 10 differential diagnosis lists generated by ChatGPT-4V was 85.1% (n=363), compared with 87.9% (n=363) by ChatGPT-4 without vision (*P*=.33). For the secondary outcome, the rate of final diagnoses as the top diagnoses generated by ChatGPT-4V was 44.4% (n=363), inferior to 55.9% (n=363) by ChatGPT-4 without vision (*P*=.002). Figure 2 shows the rate of final diagnoses within the top 10 differential diagnosis lists and as the top diagnoses generated by ChatGPT-4V and without vision.

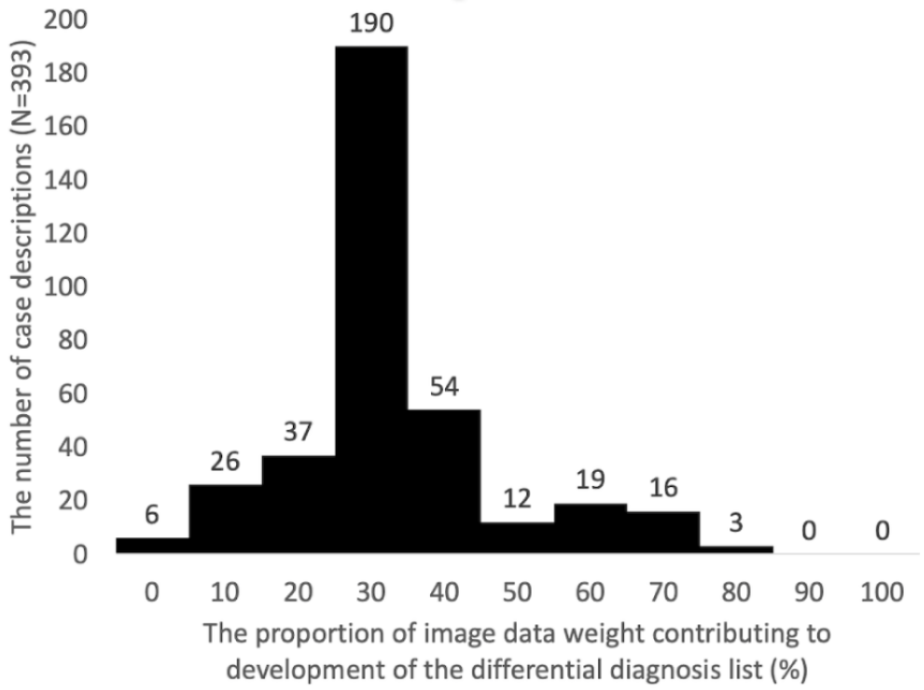
Figure 2. The rate of final diagnoses within the top 10 differential diagnosis lists and as the top diagnoses generated by ChatGPT-4 with vision and without vision.



The Contributing Weight Between Text and Image Data in Developing the Differential Diagnosis Lists

The 25th percentile, median, and 75th percentile proportions of image data weight contributing to the development of the differential diagnosis lists were 30%, 30%, and 40%, respectively, indicating a consistent reliance on image data across a significant portion of cases. The maximum and minimum proportion of image data weight contributing to the development of the differential diagnosis lists were 80% and 0%, respectively, highlighting the wide range of reliance on image data across different case reports. Specifically, in 190 case descriptions of the total 363 included case reports (190/363, 52.3%), the proportion of image data weight contributing to the development of the lists was reported to be 30%. Figure 3 shows the proportion of image data weight contributing to the development of the differential diagnosis lists.

Figure 3. The proportion of image data weight contributing to the development of the differential diagnosis lists by ChatGPT-4 with vision.



The ORs of Variables for Predicting the Outcomes

Laboratory data independently predicted the inclusion of the final diagnoses within the top 10 differential diagnosis lists by ChatGPT-4V: OR 0.52 (95% CI 0.29-0.97; *P*=.03). Additionally, MRI scans were also found to be independent predictive factors: OR 3.87 (95% CI 1.51-13.11; *P*=.01). These results were derived from univariable logistic regression models. Other variables, including the proportion of image data weight contributing to the development of the differential diagnosis lists, CT images, pathological specimens, X-ray images, and the number of characters in text data, were not associated with the final diagnoses included within the top 10 differential diagnosis lists by ChatGPT-4V, as shown in Figure 4.

Additionally, MRI scans (OR 1.93, 95% CI 1.16-3.22; *P*=.01) were independent predictive factors for the final diagnoses as top diagnoses by ChatGPT-4V. Other variables were not associated with the secondary outcome, as shown in Figure 5.

Figure 4. Odds ratios of variables for predicting the final diagnoses included within the top 10 differential diagnosis lists by ChatGPT-4 with vision in univariable regression model. *P* values are derived from the univariable logistic regression model. CT: computed tomography; MRI: magnetic resonance imaging.

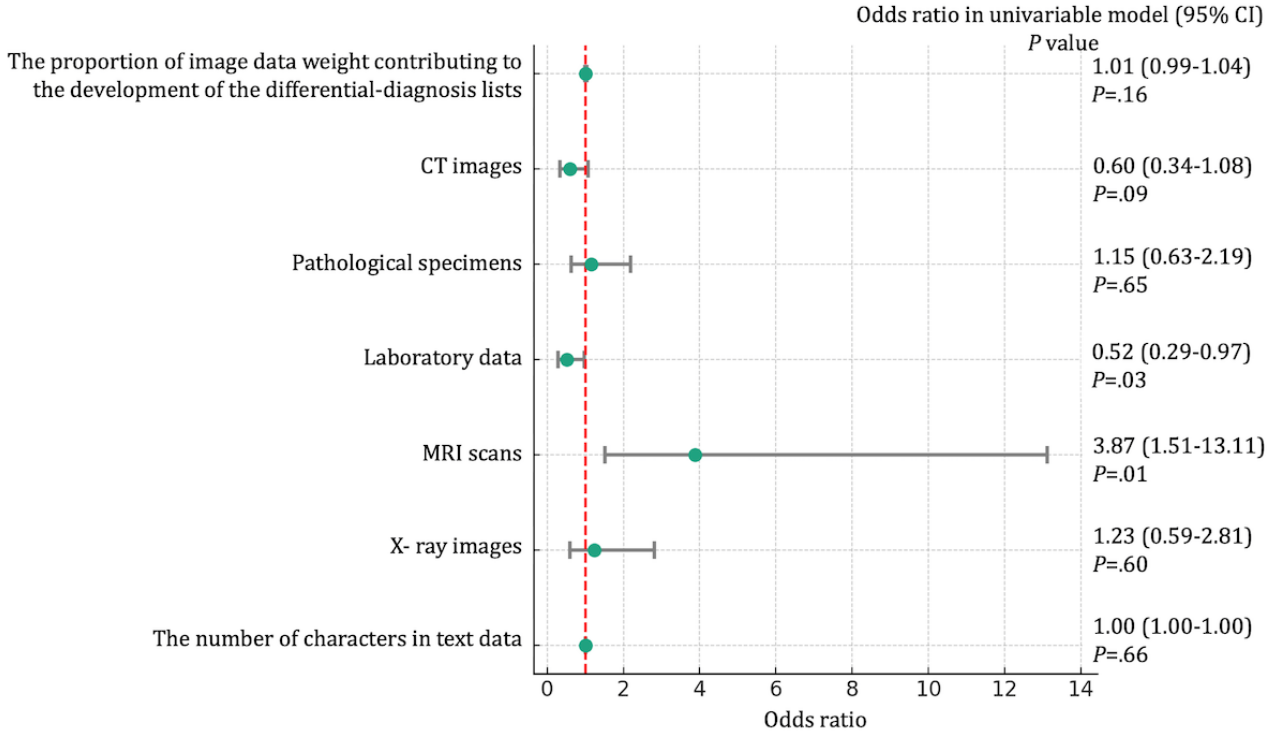
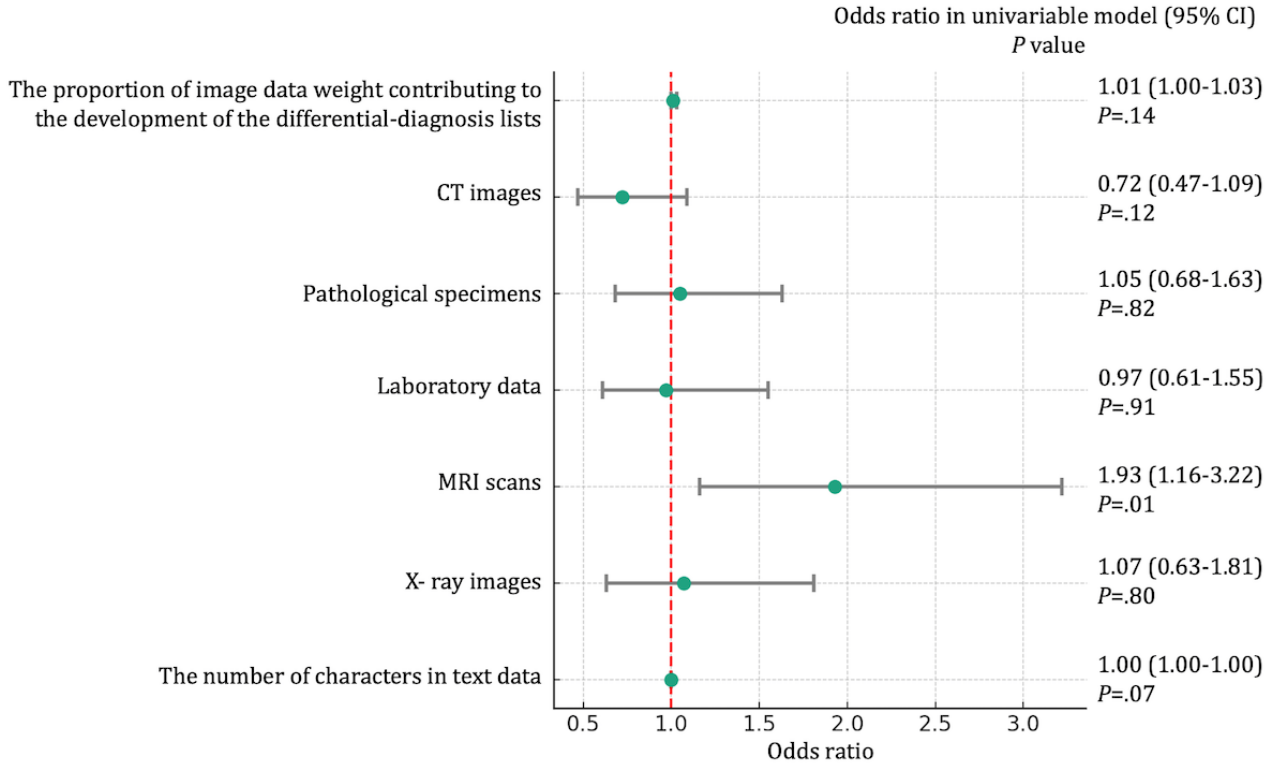


Figure 5. Odds ratios of variables for predicting the final diagnoses as top diagnoses by ChatGPT-4 with vision in univariable regression model. *P* values are derived from the univariable logistic regression model. CT: computed tomography; MRI: magnetic resonance imaging.



Discussion

Principal Results

This study showed several key findings regarding the diagnostic capabilities of ChatGPT-4 with and without vision. The incorporation of image data into ChatGPT-4V did not yield a significant improvement in diagnostic accuracy compared with that without vision. This was evident in the rates of final diagnoses within the top 10 differential diagnosis lists generated by ChatGPT-4V, where ChatGPT-4 without vision actually demonstrated comparable performance. Conversely, the rate of final diagnoses as the top diagnoses generated by ChatGPT-4V was inferior to that without vision. While ChatGPT-4V accepts a wide range of medical images, from physical examinations to various investigation results, its potential to enhance diagnostic accuracy appears underused. This underuse of image processing capabilities could be attributed to the current AI model's limitations in processing and integrating complex image data with textual data. Additionally, the AI system's training regimen, which might have emphasized text data over image data, could have resulted in a bias toward text-based analysis. Future iterations of AI systems should focus on enhancing the model's ability to discern and integrate key diagnostic features from both text and images.

In the univariable logistic regression model, these findings suggest that while the integration of image data by ChatGPT-4V did not uniformly improve diagnostic accuracy across all cases, specific types of image data, particularly MRI scans, play a crucial role in certain diagnostic contexts. MRI scans were associated with significantly higher rates of primary and secondary outcomes. Conversely, laboratory data were associated with significantly lower rates of the primary outcome. These results suggest that MRI scans are typically focused on specific body locations to target particular organs. For example, the inclusion of brain MRI scans led ChatGPT-4V to focus its differential diagnoses on cerebral diseases. The characteristics of MRI scans to focus on anatomical regions could be used to enhance the diagnostic performance of ChatGPT-4V in identifying specific conditions. Moreover, the laboratory data, often presented in tables, typically cover a broader spectrum of information than the case descriptions. For instance, in the case of infectious diseases with elevated blood glucose levels which were included only in the table, ChatGPT-4V considered hyperglycemic condition in addition to the final diagnoses. Incorporating additional laboratory data into the textual analysis could broaden the differential diagnosis lists, potentially reducing the primary outcome. The logistic regression analysis thus provides valuable insights into how different data formats influence the AI's diagnostic capabilities, guiding future improvements in AI design and training to better leverage these inputs.

Focusing on the proportion of image data weight contributing to the development of the differential diagnosis lists, a notable observation emerges regarding ChatGPT-4V's reliance. In more than half of the outputs, image data accounted for 30% of the weight in developing the differential diagnosis lists. This finding leads us to consider the system's internal decision-making process. It is important to consider that the accuracy of the proportion of image data weight in representing the actual process of integrating text and image input in ChatGPT-4V remains uncertain. Despite the consideration, the proportion of image data weight further indicates a dominant dependence on text data. It raises the possibility that ChatGPT-4V may not be

integrating text and image inputs in a balanced way. The implication here is that even with its capability to process image data, the system's diagnostic output might still be mainly influenced by text data.

Given these findings, this unexpected outcome leads us to question why additional image data did not contribute to improvements in diagnostic accuracy. Exploring the reasons behind these results, one plausible explanation emerges related to the potential biases in ChatGPT-4V's use of image data. The biases would be rooted in its training regimen. Rather than aiding in diagnosis, this image data could introduce complexity, leading ChatGPT-4V to rely more on text-based analysis and less on visual clues.

This study highlights the challenges in harnessing the full potential of multimodal AI in medical diagnostics. The findings indicate that despite the advanced capabilities of ChatGPT-4V, its integration of image data is not yet optimizing diagnostic outcomes. This would be partly because of the system's inherent design and training, which could predispose it to prioritize text over image data, despite the latter's potential richness in clinical information. This revelation is crucial for the ongoing development of AI in health care, highlighting a pivotal area for improvement. As AI continues to evolve, focusing on the harmonious integration of text and image data will be essential. This study paves the way for future innovations, guiding efforts to refine multimodal AI systems for more accurate, efficient, and reliable medical diagnostics. Future research should particularly explore the development of more sophisticated methods for image analysis and the optimization of multimodal data integration, aiming to improve the current reliance on text data and enhance the diagnostic power of AI in health care settings.

The findings from our study also raise important considerations for the practical application of AI in health care. While AI systems like ChatGPT-4V hold promise for supporting clinical decision-making, their current limitations necessitate a cautious approach to integration into clinical workflows. For instance, AI could serve as a supporting tool for preliminary analysis, helping triage or providing a second opinion in diagnostic challenges, thereby augmenting the expertise of health care professionals rather than replacing it. Health care professionals should be aware of these systems' strengths and weaknesses, leveraging them as support tools rather than definitive diagnostic solutions.

Limitations

There were several limitations in this study. A major limitation of our study was the reliance on selected image data excerpted from case reports [35], rather than whole slices of image data from clinical settings. This limitation partly arose because the current ChatGPT-4V can only process partial slices of image data [27]. This approach, while necessary for concise reporting in cases, may not accurately reflect the complexity and variability encountered in real-world clinical practice. Moreover, we excluded video files. Although generative AI systems currently do not accept video files, their inclusion could potentially improve diagnostic accuracy. Future research should explore incorporating more comprehensive image data sets and video data, technologies permitting, to enhance the AI system's diagnostic capabilities. Furthermore, the study's reliance on data derived from case reports may not encompass the diversity of real-world clinical scenarios [36]. The specificity of data sources inevitably impacts the generalizability of our findings, highlighting a significant challenge in extending our results to different health care settings and populations. Future studies should consider including complete data from real-patient scenarios with various situations.

Beyond these specific limitations, our study underscores broader concerns regarding the integration of AI in health care, particularly the potential bias inherent in the data sets used to train generative AI systems like ChatGPT-4. These biases may impact the generalizability of the AI's diagnostic and predictive capabilities across diverse populations and clinical settings. The absence of regulatory approval for generative AI systems in clinical practice further complicates their potential adoption, while inconsistencies in ChatGPT-4V interpretations of medical imaging underscore the current limitations of these technologies in performing medical functions [25].

Furthermore, the interpretability and explainability of AI-generated diagnoses remain significant hurdles [16]. The deployment of AI in health care settings also raises practical challenges related to the training of health care professionals in AI use and the integration of AI tools into existing clinical workflows. Ensuring that health care workers are adequately prepared to interpret AI-generated insights and make informed decisions is crucial for the successful adoption of AI technologies.

Last, the rapid evolution of AI technology presents unique challenges, as advancements may quickly outpace the findings of our study. The pace at which AI technologies evolve means that our conclusions may become outdated as new capabilities emerge. This highlights the importance of ongoing research and adaptation in the field of AI and health care, ensuring that studies remain relevant and that AI tools are continually evaluated and updated to reflect the latest technological advancements.

Comparison With Prior Work

Compared with a previous preliminary study for ChatGPT-4V, this study showed higher performance. The previous study assessed the proficiency of ChatGPT-4V for selected medical images from open-source libraries and repositories [27]. The study reported that only 21.7% (n=15) of cases were correctly interpreted with the correct advice. This inconsistency was partly because of the methodological differences between the 2 studies, particularly in terms of data set preparation and evaluation criteria. While the previous study mainly focused on a limited data set with simple prompts and evaluated the system's interpretation and medical advice quality, our study introduced a more comprehensive data set with a rich clinical context. Additionally, we evaluated the diagnostic accuracy, rather than merely assessing interpretation and advice, thereby providing a deeper insight into the AI system's utility in clinical decision-making.

Another study evaluated the performance of ChatGPT-4V for selected clinical cases from the website, including image data [26]. The study showed that ChatGPT-4V heavily relies on the patients' medical history. This result was consistent with this study that additional image data did not improve the diagnostic accuracy. The result was also consistent with this study that approximately half of the outputs reported that the proportion of image data weight contributing to the development of the differential diagnosis lists was 30%.

A critical distinction between our study and previous works is our comparative analysis of ChatGPT-4 with and without vision capabilities. This unique approach allowed us to highlight the impact of image data on diagnostic accuracy, revealing that while ChatGPT-4's vision component does not significantly enhance diagnostic accuracy, it does not detract from it either. This finding is crucial for understanding the role of integrated image data in AI-assisted diagnosis and highlights the potential of AI systems to support health care professionals by providing a comprehensive analysis that includes both text and image data.

Conclusions

The rates of final diagnoses within the differential diagnosis lists generated by ChatGPT-4V did not show improvement over those generated without vision. The rate of final diagnoses as the top diagnosis generated by ChatGPT-4V was inferior to that without vision. Despite its multimodal data processing capabilities, ChatGPT-4V appears to prioritize text data, which may limit its effectiveness in medical diagnostic applications, as highlighted by its system card [25]. The implications of our study for the advancement of multimodal AI systems in health care are profound. It uncovers a pivotal aspect of AI development that requires attention: the nuanced integration and weighted analysis of diverse data types. To emulate the complex reasoning of medical professionals, AI systems must advance beyond simple data incorporation toward a sophisticated synthesis that enhances diagnostic accuracy. For future improvements, we recommend the following: enhanced clinical data fusion techniques; interpretability of AI decisions; and collaborative development efforts with AI developers and medical professionals. In clinical practice, more sophisticated multimodal AI systems have the potential to enhance in providing timely, contextually rich differential diagnoses, serving as educational aids for medical trainees, and enhancing patient care by supporting remote or underserved areas. Through these enhancements, AI tools can ultimately improve patient outcomes.

Acknowledgments

TH, YH, KT, TI, 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. YH, KT, TI, 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. 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
The included cases in this study and the differential diagnosis lists generated by ChatGPT-4 with vision and without vision.
[XLSX File (Microsoft Excel File), 138 KB - medinform_v12i1e55627_app1.xlsx]

Multimedia Appendix 2
The details of image data in this study.
[DOCX File , 20 KB - medinform_v12i1e55627_app2.docx]

References

1. Yang D, Fineberg HV, Cosby K. Diagnostic excellence. JAMA 2021;326(19):1905-1906. [doi: 10.1001/jama.2021.19493] [Medline: 34709367]
2. Singh H, Connor DM, Dhaliwal G. Five strategies for clinicians to advance diagnostic excellence. BMJ 2022;376:e068044. [doi: 10.1136/bmj-2021-068044] [Medline: 35172968]
3. 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(1):17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
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;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
5. Watari T, Schiff GD. Diagnostic excellence in primary care. J Gen Fam Med 2023;24(3):143-145 [FREE Full text] [doi: 10.1002/jgf2.617] [Medline: 37261043]

6. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. Int J Environ Res Public Health 2021;18(16):8435 [FREE Full text] [doi: 10.3390/ijerph18168435] [Medline: 34444182]
7. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, The National Academies of Sciences, Engineering, and Medicine. In: Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care. Washington, DC: National Academies Press; 2015.
8. Tupasela A, Di Nucci E. Concordance as evidence in the Watson for oncology decision-support system. AI Soc 2020;35(4):811-818 [FREE Full text] [doi: 10.1007/s00146-020-00945-9]
9. Potočnik J, Foley S, Thomas E. Current and potential applications of artificial intelligence in medical imaging practice: a narrative review. J Med Imaging Radiat Sci 2023;54(2):376-385 [FREE Full text] [doi: 10.1016/j.jmir.2023.03.033] [Medline: 37062603]
10. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388(13):1201-1208. [doi: 10.1056/NEJMra2302038] [Medline: 36988595]
11. Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics 2021;22(1):122 [FREE Full text] [doi: 10.1186/s12910-021-00687-3] [Medline: 34525993]
12. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva, Switzerland: World Health Organization; 2021.
13. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568 [FREE Full text] [doi: 10.2196/48568] [Medline: 37379067]
14. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
15. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23(1):689 [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
16. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310 [FREE Full text] [doi: 10.1186/s12911-020-01332-6] [Medline: 33256715]
17. Bachute MR, Subhedar JM. Autonomous driving architectures: insights of machine learning and deep learning algorithms. Mach Learn Appl 2021;6:100164 [FREE Full text] [doi: 10.1016/j.mlwa.2021.100164]
18. Hashemi-Pour C, Kerner SM, Patrizio A. Google Gemini (formerly Bard). TechTarget. 2023. URL: https://www.techtarget.com/searchenterpriseai/definition/Google-Bard [accessed 2024-03-26]
19. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022;28(9):1773-1784 [FREE Full text] [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
20. Huang SC, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep 2020;10(1):22147 [FREE Full text] [doi: 10.1038/s41598-020-78888-w] [Medline: 33335111]
21. Jabbour S, Fouhey D, Kazerooni E, Wiens J, Sjoding MW. Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure. J Am Med Inform Assoc 2022;29(6):1060-1068 [FREE Full text] [doi: 10.1093/jamia/ocac030] [Medline: 35271711]
22. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233-1239 [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
23. OpenAI. GPT-4 technical report. ArXiv Preprint posted online on March 15 2024. [doi: 10.48550/arXiv.2303.08774]
24. ChatGPT can now see, hear, and speak. OpenAI. URL: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak [accessed 2024-03-26]
25. GPT-4V(ision) system card. OpenAI. 2023. URL: https://openai.com/research/gpt-4v-system-card [accessed 2024-03-26]
26. Wu C, Lei J, Zheng Q, Zhao W, Lin W, Zhang X, et al. Can GPT-4V(ision) serve medical applications? case studies on GPT-4V for multimodal medical diagnosis. ArXiv Preprint posted online on December 04, 2023. [doi: 10.48550/arXiv.2310.09909]
27. Senkaiahliyan S, Toma A, Ma J, Chan AW, Ha A, An KR, et al. GPT-4V(ision) unsuitable for clinical care and education: a clinician-evaluated assessment. medRxiv Preprint posted online on November 16, 2023. [doi: 10.1101/2023.11.15.23298575]
28. Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in Japanese national medical licensing examination. medRxiv Preprint posted online on November 08, 2023. [doi: 10.1101/2023.11.07.23298133]
29. Driessen T, Dodou D, Bazilinskyy P, de Winter J. Putting ChatGPT Vision (GPT-4V) to the test: risk perception in traffic images. ResearchGate. 2023. URL: https://www.researchgate.net/publication/375238184_Putting_ChatGPT_Vision_GPT-4V_to_the_test_Risk_perception_in_traffic_images [accessed 2024-03-26]
30. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision). ArXiv Preprint posted online on October 11, 2023. [doi: 10.48550/arXiv.2309.17421]
31. Yang J, Zhang H, Li F, Zou X, Li C, Gao J. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. ArXiv Preprint posted online on November 06, 2023. [doi: 10.48550/arXiv.2310.11441]

32. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 2008;23(Suppl 1):37-40 [FREE Full text] [doi: 10.1007/s11606-007-0271-8] [Medline: 18095042]
33. Miao J, Gibson LE, Craici IM. Levofloxacin-associated bullous pemphigoid in a hemodialysis patient after kidney transplant failure. Am J Case Rep 2022;23:e938476 [FREE Full text] [doi: 10.12659/AJCR.938476] [Medline: 36578185]
34. Krupat E, Wormwood J, Schwartzstein RM, Richards JB. Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors. Med Educ 2017;51(11):1127-1137. [doi: 10.1111/medu.13382] [Medline: 28857266]
35. Riley DS, Barber MS, Kienle GS, Aronson JK, von Schoen-Angerer T, Tugwell P, et al. CARE guidelines for case reports: explanation and elaboration document. J Clin Epidemiol 2017;89:218-235 [FREE Full text] [doi: 10.1016/j.jclinepi.2017.04.026] [Medline: 28529185]
36. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. J Med Internet Res 2022;24(10):e37408 [FREE Full text] [doi: 10.2196/37408] [Medline: 36287594]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
ChatGPT-4V: ChatGPT-4 with vision
CT: computed tomography
LLM: large language model
MRI: magnetic resonance imaging
OR: odds ratio

Edited by A Castonguay; submitted 18.12.23; peer-reviewed by D Hu, D Singh, TAR Sure; comments to author 07.02.24; revised version received 14.02.24; accepted 13.03.24; published 09.04.24.

Please cite as:
Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T
Evaluating ChatGPT-4's Diagnostic Accuracy: Impact of Visual Data Integration
JMIR Med Inform 2024;12:e55627
URL: https://medinform.jmir.org/2024/1/e55627
doi:10.2196/55627
PMID:38592758

Original Paper

Natural Language Processing–Powered Real-Time Monitoring Solution for Vaccine Sentiments and Hesitancy on Social Media: System Development and Validation

Liang-Chin Huang[1], PhD; Amanda L Eiden[2], PhD; Long He[1], MS; Augustine Annan[1], PhD; Siwei Wang[1], MS; Jingqi Wang[1], PhD; Frank J Manion[1], PhD; Xiaoyan Wang[1], PhD; Jingcheng Du[1], PhD; Lixia Yao[2], PhD

[1]Melax Tech, Houston, TX, United States

[2]Merck & Co, Inc, Rahway, NJ, United States

Corresponding Author:
Amanda L Eiden, PhD
Merck & Co, Inc
2025 E Scott Ave
Rahway, NJ, 07065
United States
Phone: 1 7325944000
Email: amanda.eiden@merck.com

Abstract

Background: Vaccines serve as a crucial public health tool, although vaccine hesitancy continues to pose a significant threat to full vaccine uptake and, consequently, community health. Understanding and tracking vaccine hesitancy is essential for effective public health interventions; however, traditional survey methods present various limitations.

Objective: This study aimed to create a real-time, natural language processing (NLP)–based tool to assess vaccine sentiment and hesitancy across 3 prominent social media platforms.

Methods: We mined and curated discussions in English from Twitter (subsequently rebranded as X), Reddit, and YouTube social media platforms posted between January 1, 2011, and October 31, 2021, concerning human papillomavirus; measles, mumps, and rubella; and unspecified vaccines. We tested multiple NLP algorithms to classify vaccine sentiment into positive, neutral, or negative and to classify vaccine hesitancy using the World Health Organization's (WHO) 3Cs (confidence, complacency, and convenience) hesitancy model, conceptualizing an online dashboard to illustrate and contextualize trends.

Results: We compiled over 86 million discussions. Our top-performing NLP models displayed accuracies ranging from 0.51 to 0.78 for sentiment classification and from 0.69 to 0.91 for hesitancy classification. Explorative analysis on our platform highlighted variations in online activity about vaccine sentiment and hesitancy, suggesting unique patterns for different vaccines.

Conclusions: Our innovative system performs real-time analysis of sentiment and hesitancy on 3 vaccine topics across major social networks, providing crucial trend insights to assist campaigns aimed at enhancing vaccine uptake and public health.

(JMIR Med Inform 2024;12:e57164) doi:10.2196/57164

KEYWORDS

vaccine sentiment; vaccine hesitancy; natural language processing; NLP; social media; social media platforms; real-time tracking; vaccine; vaccines; sentiment; sentiments; vaccination; vaccinations; hesitancy; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; machine learning; uptake; willing; willingness; classification

Introduction

Vaccine is an essential public health intervention that has saved millions of lives and achieved a substantial global reduction in cases, hospitalizations, and health care costs associated with vaccine-preventable diseases (VPDs) [1-3]. Yet, despite their value, vaccine hesitancy persists as a barrier to full vaccine uptake. The World Health Organization (WHO) defines vaccine hesitancy as the delay or refusal of vaccination, even when vaccination services are accessible [4]. Additionally, the WHO identifies vaccine hesitancy as one of the top 10 global health threats [5]. Delay or refusal of vaccines due to vaccine hesitancy can have broad-reaching implications; unvaccinated individuals not only put themselves at risk of VPDs, such as COVID-19,

but also pose a threat to the broader community or even global health [6]. This phenomenon has been documented since the advent of vaccines in over 90% of the countries [7]. Considering the case of measles, mumps, and rubella (MMR), it is crucial to uphold community protection or herd immunity, necessitating widespread vaccination to protect those unable to receive the vaccine [8]. A former London study successfully raised MMR vaccination rates from 80% to 94% in under 2 years through incentivized care packages and innovative technology use, approaching the desired herd immunity target [9].

There is a myriad of reasons for vaccine hesitancy, including personal or familial beliefs, concerns about adverse reactions or efficacy, and skepticism toward government and vaccine manufacturers [6,10-17]. This intricate web of motivations makes vaccine hesitancy a complex public health challenge [18].

Understanding vaccine hesitancy is crucial for developing effective interventions, public health education, and vaccination promotion strategies [19-22]. While surveys have traditionally served as a valuable tool for gathering public opinions on vaccination, they possess inherent limitations such as static data collection, resource intensiveness, and potential time lag [23-29]. To address these limitations, real-time tracking of vaccine hesitancy activities and trends offers public health professionals' valuable insights. This approach helps identify critical intervention points before the vaccination uptake wanes, allowing for more targeted and timely communication efforts.

The emergence of social media platforms has enabled billions of users to engage in discussions, information sharing, and opinion expression on various subjects, including health-related topics [30]. While this presents an unprecedented opportunity for public health improvement, it also poses a significant risk linked to the dissemination of vaccine-related misinformation and disinformation [31]. Previous research has used semiautomatic methods such as manual coding and hashtag or keyword analysis to study social media vaccine discussions [32-35]. Nevertheless, these approaches may sometimes encounter potential challenges with scalability and precision. Natural language processing (NLP) is an automated method designed to effectively and accurately decipher the wealth of information in natural language text, addressing challenges such as ambiguities and probabilistic parsing, and enabling applications such as information extraction and discourse analysis [36]. This technique has emerged as a promising solution, holding the potential to mitigate these challenges and improve the precision of vaccine-related public sentiment analysis [37,38].

To address these challenges, this study's principal aim was to create an NLP system for real-time monitoring of vaccine sentiment and hesitancy across English-language social media platforms targeting the US market. Our 3-fold contributions are (1) developing one of the first real-time monitoring systems for social media vaccine discussions that covers 3 major social media platforms and 3 vaccine topic groups [39]; (2) comprehensively evaluating multiple machine learning–based NLP models for social media post classification tasks, thus establishing a benchmark for future research; and (3) analyzing decade-long trends of sentiment and hesitancy and linked real-world events to corresponding points on the trends for multiple vaccine targets.

Methods

Overview

We followed a systematic approach to monitor vaccine sentiment and hesitancy posts on Twitter (subsequently rebranded as X), Reddit, and YouTube. We selected Twitter, Reddit, and YouTube as they are the primary social media platforms offering substantial volumes of posts through application programming interface (API) access [40-43]. We focused exclusively on English language posts given the widespread use of English in the largest market countries for our target vaccines and with regard to the accessibility of English language social media. Other platforms and languages, such as Facebook and Spanish [44], may be of interest for future studies; however, these served as a first approach to research. Figure 1 illustrates our workflow, including data annotation, NLP algorithms, and an online dashboard.

First, we categorized vaccine sentiment into positive, negative, and neutral, which were the labels also used in other sentiment analyses using social media data [45,46]. Then, we aligned vaccine hesitancy with the WHO's 3Cs (confidence, complacency, and convenience) vaccine hesitancy model, described in further detail in the *3Cs Vaccine Hesitancy Annotation* section [4]. The definitions of post sentiment and vaccine hesitancy are comprehensively presented in Table 1. We collected data using vaccine-specific search queries (see Table S1 in Multimedia Appendix 1) for relevant posts from the 3 social media platforms. To ensure the quality and reliability of the data, we collaborated with medical experts to create annotated corpora aligned with the information model. These corpora were then used to train NLP algorithms to automatically extract vaccine sentiment and hesitancy content. Finally, we developed an online dashboard to provide real-time insights into vaccine sentiment and hesitancy trends. Our study focuses on evaluating the vaccine sentiment and hesitancy of human papillomavirus (HPV), MMR, and general or unspecified vaccines. The critical role of the vaccines is exemplified by the HPV vaccine, which has effectively reduced prevalent HPV infections and precancerous lesions, underlining the importance of global implementation [47], and the MMR vaccine is renowned for its safety and efficacy, which has greatly mitigated endemic diseases in the United States [48]. Despite these successes, challenges such as insufficient vaccination coverage, increasing hesitancy, and the resurgence of mumps, attributed to waning immunity and antigenic variation, persist worldwide. Throughout the COVID-19 pandemic up to 2022, HPV and MMR were the vaccines that maintained the greatest negative impact on routine vaccinations in the United States, suggesting a need for proactive efforts to increase vaccination coverage to prevent associated health complications and costs [49].

Figure 1. The overview of study design and classifications used to evaluate vaccine-related posts. 3Cs: confidence, complacency, and convenience; ML: machine learning; WHO: World Health Organization.

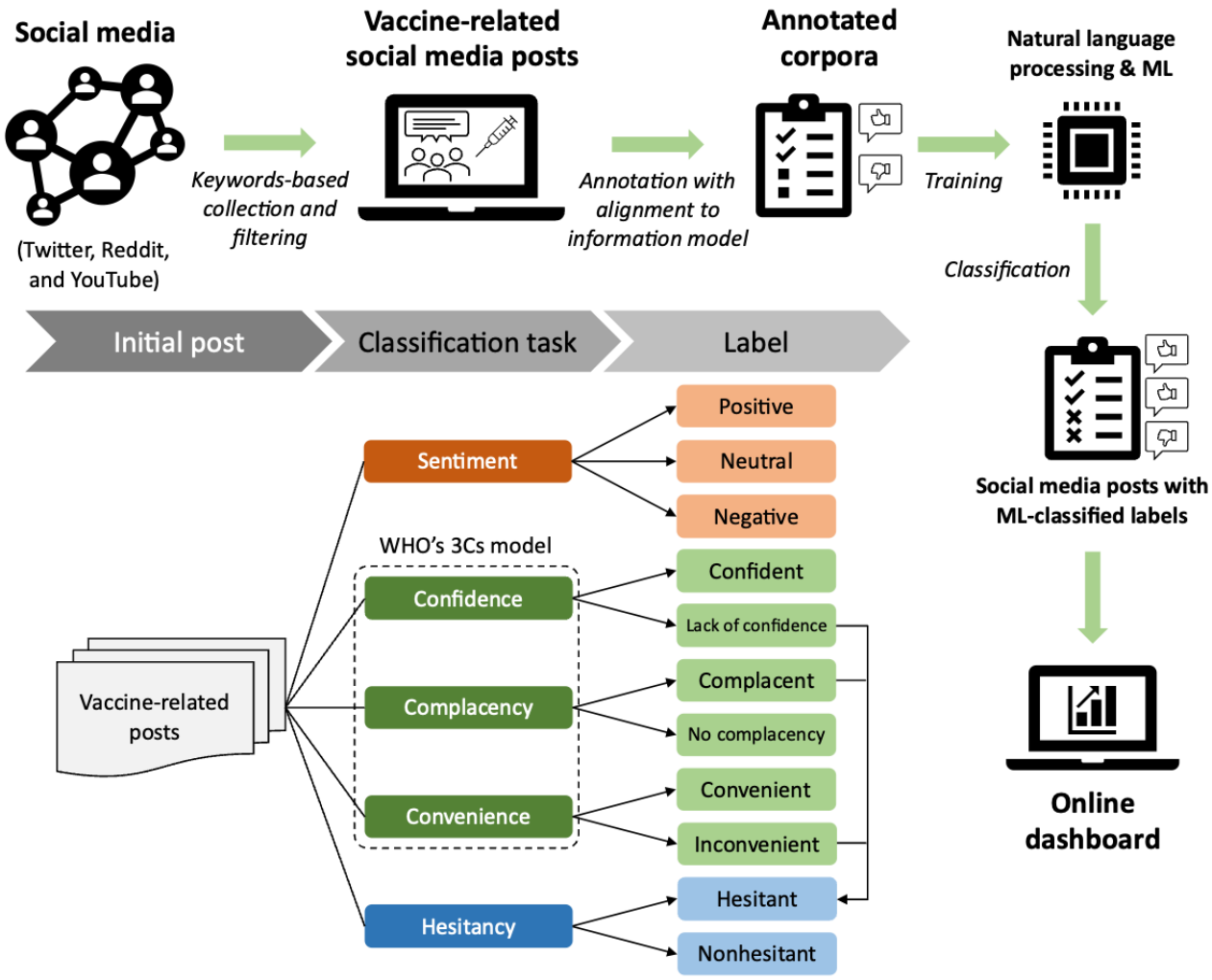


Table 1. Definitions of post sentiment and hesitancy.

Classification task and label	Definition
Sentiment	
Positive	Posts that mention, report, or share positive news, opinions, or stories about vaccines or vaccination.
Neutral	Posts that are related to vaccines or vaccination topics but contain no sentiment, the sentiment is unclear, or they contain both negative and positive sentiments.
Negative	Posts that mention, report, or share negative news, opinions, or stories about vaccines or vaccination, which may discourage vaccination.
Confidence	
Confident	Posts reflecting a trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers' motivations.
Lack of confidence	Posts reflecting a lack of trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers' motivations.
Complacency	
Complacent	Posts where the perceived risks of VPDs[a] are low and vaccination is deemed as an unnecessary preventive action.
No complacency	Posts where the perceived risks of VPDs are high and vaccination is deemed as a necessary preventive action.
Convenience	
Convenient	Posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services do not affect uptake.
Inconvenient	Posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services affect uptake.
Hesitancy	
Hesitant	The post is labeled as lack of confidence, complacent, or inconvenient.
Nonhesitant	The post is not labeled as lack of confidence, complacent, or inconvenient.

[a]VPD: vaccine-preventable disease.

Social Media Data Collection

The systematic collection of social media data spanned from January 1, 2011, to October 31, 2021, across 3 platforms—Twitter, Reddit, and YouTube. During initial exploratory analysis, we recognized variations in text nature and query logic across these platforms, leading us to tailor our search queries for each platform to collect relevant posts while excluding irrelevant ones. Table S1 in Multimedia Appendix 1 lists the customized queries on each platform for each vaccine topic group, which include both inclusion and exclusion keywords. We retrieved the results (relevant posts) using the APIs provided by the 3 platforms. Details about the software versions are described in the Multimedia Appendix 1. To clarify ethical considerations and data privacy issues, when gathering data from Twitter, YouTube, and Reddit, we adhered to their API's data privacy policies and ensured the deidentification of all posts and videos by assigning them a unique random ID.

Ethical Considerations

Ethics board review was not required, as all modelling data came from public sources and there were no ethical issues. The data privacy policies of the application program interfaces (APIs) of Twitter, YouTube, and Reddit were followed when gathering data. We ensured the deidentification of all posts and videos by assigning them a unique random ID.

Data Annotation

From the retrieved results, approximately 90 million posts, we randomly selected 60,000 social media discussions. These posts were manually annotated to build both training and evaluation data sets, which were used for building the text classifiers. We selected 20,000 posts for annotation, including 10,000 tweets, 5000 Reddit posts, and 5000 YouTube comments for each vaccine topic group, including HPV vaccine, MMR vaccine, and general or unspecified vaccines. During annotator training, 4 annotators with a medical training background were recruited for the annotation. An annotation guideline was developed. All annotators first annotated the same 1000 tweets, 1000 Reddit posts, and 1000 YouTube posts independently, and then discussed collectively for any discrepancies. After all discrepancies were resolved through discussions, these annotators began to annotate the rest of the social media posts. A 2-fold annotation strategy was used, where first, we annotated the sentiment of the post as positive, neutral, or negative, assigning only 1 category to each post; and second, we annotated vaccine hesitancy based on the constructs of the WHO 3Cs model, which include confidence, complacency, and convenience (Figure 1). These annotation categories also define each classification task.

Sentiment Annotation

The annotation task involved assigning 1 of 3 sentiment labels to each post, which constituted a multiple-class classification problem. The labels and corresponding illustrative examples are defined in Textbox 1.

Textbox 1. Definitions and examples of sentiment labels.

- Positive: posts that mention, report, or share positive news, opinions, or stories about vaccines or vaccination.
  - Example: "HPV vaccine, prevents against the two HPV types, 16 and 18, which cause 70% of cervical cancers"
  - Example: "Get vaccinated against HPV to protect you in the future for now!"
- Neutral: posts that are related to vaccines or vaccination topics but contain no sentiment, the sentiment is unclear, or they contain both negative and positive sentiments.
  - Example: "The following report is specifically for the MMR vaccine, but you can browse around for others"
  - Example: "I just learned that there are more than 50 strains of HPV...I always thought the vaccine prevented all strains."
- Negative: posts that mention, report, or share negative news, opinions, or stories about vaccines or vaccination, which may discourage vaccination.
  - Example: "According to a report, thousands of kids suffer permanent injury or death by getting vaccines"
  - Example: "Believe it? Vaccines have killed 1000 more kids than any measles!"

3Cs Vaccine Hesitancy Annotation

The annotation task involved assigning multiple labels to each post according to the 3Cs model constructs. Annotators checked each construct to determine whether the post was related to it separately. If any of the constructs were labeled as "lack of confidence," "complacent," or "inconvenient," we considered the post as vaccine hesitant; otherwise, it was considered vaccine nonhesitant. Definitions and examples for each 3Cs model construct are provided in Textbox 2.

Table S2 in Multimedia Appendix 1 provides examples of specific social media posts with annotations for the different categories. The distribution of annotated posts in each sentiment and 3Cs construct for each platform and vaccine topic group is shown in Table S3 in Multimedia Appendix 1.

Textbox 2. Definitions and examples of World Health Organization's 3Cs (confidence, complacency, and convenience) model.

- Lack of confidence: posts reflecting a lack of trust in the effectiveness and safety of vaccines, the vaccine delivery system, or policy makers' motivations.
  - Example: "Fully vaccinated are 30 times more likely to get COVID-19, and 10 times more likely to require hospitalization."
  - Example: "The vaccine label includes all these events. Concerns have been raised about reports of deaths occurring in individuals after receiving that vaccine."
- Complacency: posts where the perceived risks of vaccine-preventable diseases are low, and vaccination is deemed as an unnecessary preventive action.
  - Example: "Why do adults need to know about the measles vaccine? The measles is a benign disease and there is no need for vaccines."
  - Example: "I wasn't vaccinated against a preventable disease. It's not always just a life-or-death dichotomy - I recovered."
- Inconvenience or convenience: posts where physical availability, affordability and willingness to pay, geographical accessibility, ability to understand (language and health literacy), and appeal of immunization services affect uptake.
- Example: "I am 30-year-old man and am looking for an HPV vaccine. Unfortunately, my insurance only covers it for women. I am particularly at risk for certain cancers. I really don't understand how insurance companies are allowed to make the gender distinction when the FDA approved it for both."

Text Classification Algorithms

Overview

To classify the sentiment and hesitancy of social media posts, we compared the performance of 5 text classification algorithms—logistic regression (LR) [50], support vector machine (SVM) [51], random forest [52], extreme gradient boosting (XGBoost) [53], and Snorkel [54]. Each of these models has unique characteristics, which are summarized below.

LR Algorithm

LR is a classic statistical methodology that models a binary dependent variable using a logistic function. It is favored in medical research due to its ability to determine the odds ratio, indicating the potential change in outcome probabilities [55].

SVM Algorithm

SVM is one of the most robust classification methods based on statistical learning frameworks. It finds a hyperplane in an N-dimensional space that distinctly classifies data points. In

medical text mining, SVM combined with other algorithms has demonstrated effective performance in extracting and recognizing entities in clinical text, contributing notably to improved patient care [56].

Random Forest Algorithm

Random forest is a classifier that uses ensemble learning to combine decision tree classifiers through bagging or bootstrap aggregating. It has been applied to highly ranked features obtained through suitable ranker algorithms and has shown promising results in medical data classification tasks, enhancing the prediction accuracy for various diseases [57].

XGBoost Algorithm

XGBoost is an ensemble of algorithms that turn weak learners into strong learners by focusing on where the individual models went wrong. In gradient boosting, individual weak models train upon the difference between the classification and the actual results. It has been effective in mining and classifying suggestive sentences from online customer reviews by combining them with a word-embedding approach [58].

Snorkel Algorithm

Snorkel is a system that enables users to train models without hand labeling all training data by writing their labeling functions. Using Snorkel enables the extraction of chemical reaction relationships from biomedical literature abstracts, supporting the understanding of biological processes without requiring a large, labeled training data set [59].

We extracted the term frequency–inverse document frequency vector for each word in all text classification algorithms using *scikit-learn*'s *TfidfTransformer* function with default parameter settings. Term frequency–inverse document frequency evaluates how relevant a word is to a text in a collection of texts [60]. If the model encounters a new post with words or symbols not included in its original bag of words, it will effectively ignore those words during the transformation process. To ensure a balanced training set, the 3 class-balancing methods implemented by Python *imblearn* package applied were (1) random oversampling, (2) synthetic minority over-sampling technique (SMOTE) [61], and (3) SVM-based SMOTE [62] (with the default parameter settings, specifically k_neighbors=5, as they exhibited the optimal performance within the developer's data set [61]). SMOTE randomly selects a minority class instance, finds one of its nearest minority class neighbors, and then synthesizes an instance between these 2 instances in the feature space. SVM-based SMOTE uses support vectors to determine the decision boundaries and then synthesizes a minority class instance along the decision boundary.

NLP Evaluation

The evaluation data sets were created from the annotated corpora and randomly divided into training, validation, and test sets in a 6:2:2 ratio to assess the performance of the 5 text classification algorithms. The models were trained on the training sets, optimized on the validation sets, and then evaluated on the test sets. The following key metrics were calculated to evaluate the models:

A true positive occurs when the model accurately classifies the positive class (positive, negative, or neutral for sentiment; true for 3Cs model constructs). A true negative occurs when the model accurately classifies the negative class (nonpositive, nonnegative, or nonneutral for sentiment; false for 3Cs model constructs). A false positive is an incorrect positive classification, while a false negative is an incorrect negative classification. As the sentiment and hesitancy labels in Tweets, Reddit posts, and YouTube comments are imbalanced, we optimized our models based on F_1-scores, which balance precision and recall, rather than accuracy. The purpose of optimizing a model based on F_1-scores when dealing with imbalanced labels is to achieve a better balance between precision and recall, thereby improving the overall performance of the model. This is especially important in imbalanced data sets where the cost of misclassification can be high.

Dashboard Development

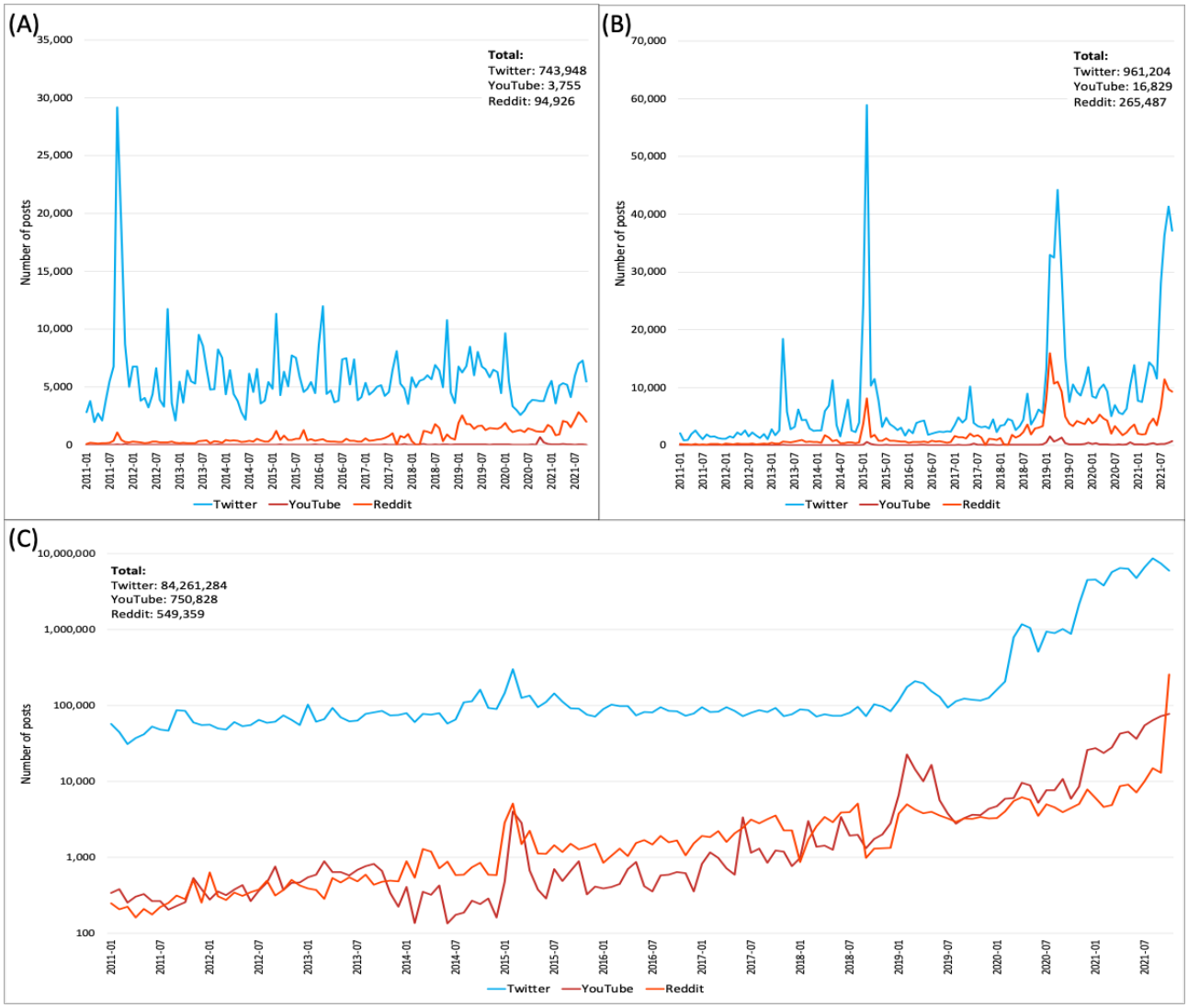
We designed a user-friendly, web-based visualization dashboard [39] for real-time analysis of trends in vaccine sentiment and hesitancy over time and geography (Figure S1A-C in Multimedia Appendix 1). The dashboard also allows for comparisons of sentiment and hesitancy across different social media platforms and vaccine topic groups (Figure S1D in Multimedia Appendix 1). The NLP models were optimized based on their F_1-scores to address the imbalanced labels of sentiment and hesitancy in tweets, Reddit posts, and YouTube comments. The selected models are applied to all unlabeled data collected from 2011 to 2021. Technical details are described and represented in Figure S2 in Multimedia Appendix 1.

Results

Social Media Data Collection Summary

From January 1, 2011, to October 31, 2021, we collected 86 million posts from Twitter, 0.9 million from Reddit, and 76,000 from YouTube, which were related to vaccines. The most widely discussed topic across all 3 platforms was the general or unspecified vaccine, followed by the MMR and then HPV vaccines. We observed a substantial increase in the general vaccine-related discussions on Twitter and Reddit starting in early 2020, coinciding with the onset of the COVID-19 pandemic. The collected social media data and growth trends are plotted in Figure 2.

Figure 2. The long and short-term trends of collected vaccine-related social media post data across 3 different platforms for different vaccine topic groups: (A) human papillomavirus (HPV); (B) measles, mumps, and rubella (MMR); and (C) general or unspecified vaccine.



NLP Performance on Vaccine Sentiment and Hesitancy

We tested all combinations of the 5 NLP algorithms. The performances in sentiment classification, hesitancy classification, and 3Cs classifications are presented in Table 2. The best-performing algorithms (according to F_1-scores) and detailed performance scores for different classification tasks are shown in Tables S4-S8 in Multimedia Appendix 1. In sentiment classification, LR outperformed other algorithms in 7 out of 9 platform–vaccine topic group combinations, with overall accuracies ranging from 0.51 to 0.78 (Table S4 in Multimedia Appendix 1). The macroaveraged F_1-scores of negative, neutral, and positive sentiment classifications across different platforms and vaccine topic groups were 0.43, 0.67, and 0.53, respectively. In hesitancy classification, LR outperformed other algorithms in 6 platform–vaccine topic group combinations, with overall accuracies ranging from 0.69 to 0.91 (Table S5 in Multimedia Appendix 1). The macroaveraged F_1-scores of nonhesitancy and hesitancy classifications were 0.86 and 0.40, respectively. Notably, Reddit users had fewer negative sentiment posts, resulting in lower performance in classifying negative sentiment. In addition, as Reddit had fewer hesitancy posts, classifying hesitancy was more challenging than on Twitter and YouTube.

Our evaluation of various algorithms and class-balancing methods for each platform and vaccine topic group revealed that Snorkel performed best in 3 platform–vaccine topic group combinations in vaccine hesitancy classifications, with overall accuracies ranging from 0.69 to 0.98 (Table S6 in Multimedia Appendix 1). The macroaveraged F_1-scores for lack of confidence and nonlack of confidence classifications were 0.88 and 0.45, respectively. Similarly, for complacency classifications, Snorkel outperformed other algorithms in 4 platform–vaccine topic group combinations, with overall accuracies ranging from 0.64 to 0.99 (Table S7 in Multimedia Appendix 1). The macroaveraged F_1-scores for noncomplacency and complacency classifications were 0.89 and 0.49, respectively. Inconvenience classifications were significantly improved with Snorkel in 8 platform–vaccine topic group combinations, with overall accuracies ranging from 0.89 to 0.99 (Table S8 in Multimedia Appendix 1). However, the results are biased as there were limited posts with convenience information on all 3 social media platforms, which may impact generalizability. The macroaveraged F_1-scores for

noninconvenience and inconvenience classifications were 0.98 and 0.38, respectively. Our findings demonstrate that advanced text classification algorithms such as XGBoost and Snorkel outperformed other algorithms in highly class-imbalanced situations, even when different class-balancing methods were applied.

We have created a web-based dashboard building upon those best-performing NLP algorithms to extract vaccine sentiment and hesitancy from social media posts. The dashboard summarizes posts from the 3 social media platforms and allows users to analyze temporal trends and geographic clustering easily. It offers different views, including 3 social media platform–centric views and a comparison view that enables users to compare selected vaccine topic groups and sentiment or hesitancy (Figure S1 in Multimedia Appendix 1).

When analyzing the sentiment of HPV vaccine posts across 3 social media platforms from January 2011 to October 2021 (Figure 3A), we observed that the ratio of positive sentiment was generally higher than that of neutral and negative sentiment. We also compared vaccine sentiment across 3 social media platforms for MMR vaccines from January 2011 to October 2021 (Figure 3B). Overall, posts expressed positive sentiment toward MMR, with most being neutral. Taking the hesitancy of MMR vaccine as an example, the overall trend shows that the social media posts across 3 social media platforms have a higher ratio of nonhesitancy than hesitancy (Figure 3C).

Table 2. NLP[a] performance (measured by F1-scores and accuracy) on vaccine sentiment and hesitancy.

Performance	Twitter			Reddit			YouTube		
	HPV[b]	MMR[c]	General[d]	HPV	MMR	General	HPV	MMR	General
Sentiment									
Positive F_1-score	0.87	0.57	0.47	0.67	0.50	0.35	0.58	0.53	0.19
Neutral F_1-score	0.71	0.67	0.83	0.67	0.65	0.86	0.51	0.59	0.51
Negative F_1-score	0.41	0.53	0.43	0.32	0.26	0.21	0.60	0.49	0.59
Accuracy	0.78	0.61	0.73	0.63	0.55	0.75	0.56	0.55	0.51
Confidence									
Confident F_1-score	0.35	0.31	0.52	0.35	0.62	0.56	0.44	0.29	0.63
Lack of confidence F_1-score	0.88	0.95	0.79	0.86	0.74	0.84	0.89	0.99	0.98
Accuracy	0.80	0.90	0.71	0.77	0.69	0.77	0.82	0.98	0.95
Complacency									
Complacent F_1-score	0.47	0.36	0.41	0.43	0.68	0.60	0.33	0.50	0.60
No complacency F_1-score	0.94	0.91	0.81	0.93	0.59	0.96	0.91	1.00	0.97
Accuracy	0.89	0.84	0.71	0.88	0.64	0.93	0.84	0.99	0.95
Convenience									
Convenient F_1-score	0.96	0.99	0.99	0.94	0.95	0.98	0.98	1.00	0.99
Inconvenient F_1-score	0.48	0.18	0.55	0.67	0.17	0.50	0.17	0.50	0.20
Accuracy	0.92	0.98	0.98	0.89	0.91	0.97	0.96	0.99	0.98
Hesitancy									
Hesitant F_1-score	0.40	0.44	0.38	0.19	0.23	0.20	0.58	0.53	0.61
Nonhesitant F_1-score	0.94	0.90	0.89	0.87	0.81	0.95	0.81	0.83	0.76
Accuracy	0.90	0.83	0.82	0.78	0.69	0.91	0.73	0.75	0.70

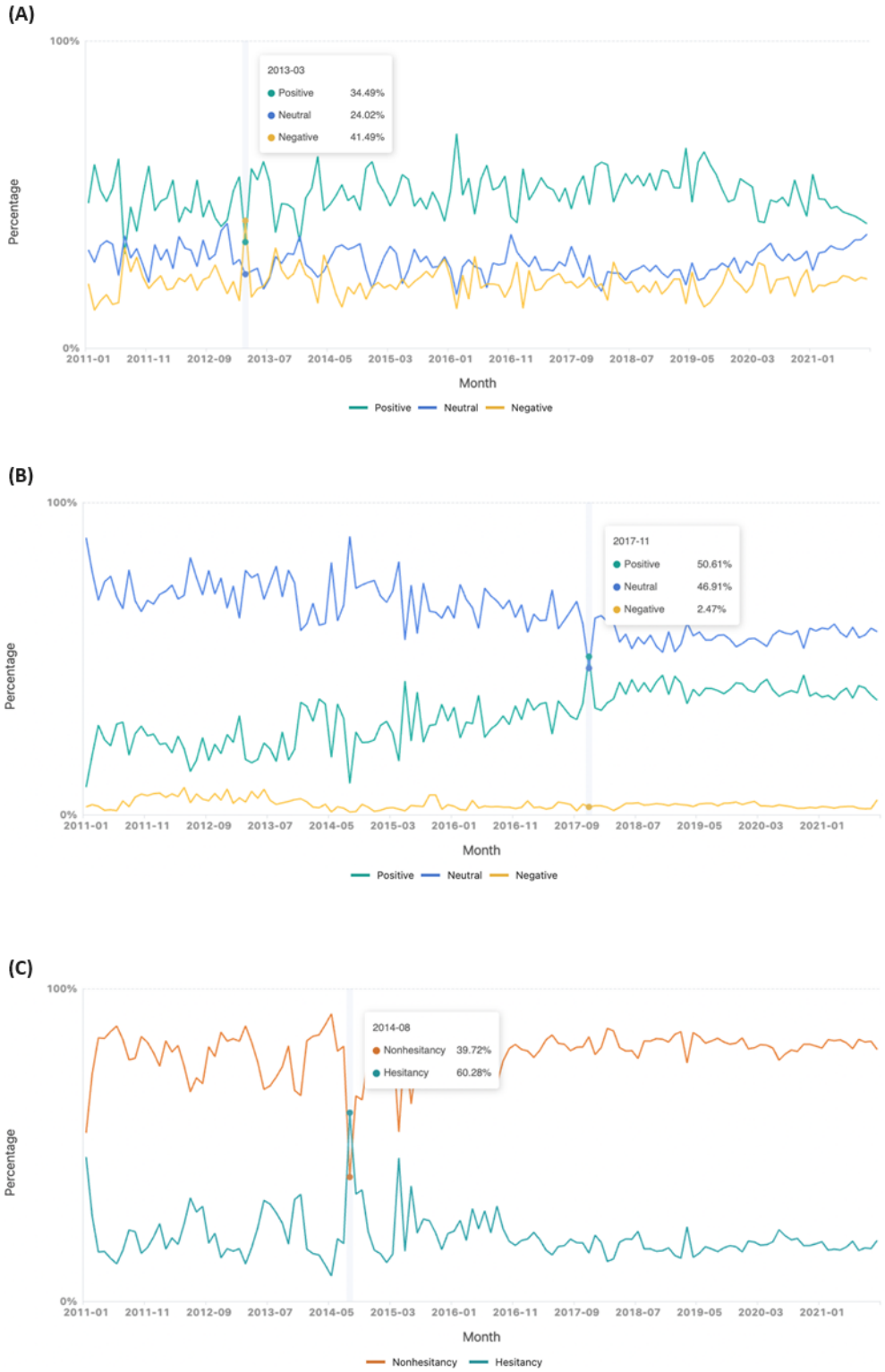
[a]NLP: natural language processing.

[b]HPV: human papillomavirus.

[c]MMR: measles, mumps, and rubella.

[d]General: general or unspecified vaccines.

Figure 3. Temporal trends of vaccine sentiment and hesitancy. (A) Aggregation of 3 social media platform data sources to evaluate vaccine sentiment for HPV vaccine–related posts. (B) Comparison of vaccine sentiment for MMR vaccines. (C) Comparison of vaccine hesitancy for MMR vaccine. HPV: human papillomavirus; MMR: measles, mumps, and rubella.



Discussion

Principal Findings

Our analysis of temporal trends in vaccine-related sentiment on social media platforms yielded valuable insights into the dynamics of public perception. A total of 5 different classification algorithms were subjected to tests for performance in sentiment and hesitancy classifications, revealing that advanced text classification algorithms such as XGBoost and Snorkel outperformed others in classifying hesitancy, complacency, and other factors, while LR had a superior performance for sentiment classification. The superior performance of LR could potentially be attributed to its enhanced ability to effectively handle binary classification

challenges and manage noise variables [63]. As the use of artificial intelligence platforms is increasingly becoming accessible for public use, it is crucial to gain an understanding of their accuracy and limitations. Traditional machine learning algorithms have the ability to predict outcomes but often lack transparency. Hence, enhancing public understanding and advancing toward explainable artificial intelligence is vital for error rectification and improved model efficacy for social media research [64].

When evaluating trends for the HPV vaccine, overall positive sentiment outweighed neutral and negative sentiment (Figure 3A), a notable exception occurred in March 2013. During this period, posts with negative sentiment on all 3 platforms surpassed those with 34% (2270/6582) positive and 24% (1581/6582) neutral sentiment, constituting 41% (2731/6582) of the total. This spike in negative sentiment can be attributed to news articles published in March 2013; for example, "Worried Parents Balk At HPV Vaccine For Daughters" by National Public Radio [65] and "Side Effect Fears Stop Parents from Getting HPV Vaccine for Daughters" by CBS News [66]. These articles highlighted concerns and fears about the HPV vaccine. Afterward, specific studies were conducted and published to further investigate these concerns and fears [67,68]. Notably, the HPV vaccines have been found to be safe in several studies and strongly recommended by the Centers for Disease Control and Prevention (CDC), etc [69,70].

Conversely, overall, posts expressed more neutral sentiment toward MMR than positive sentiment (Figure 3B), with an exception in November 2017. During this month, 51% (2844/5619) of posts expressed positive sentiment and 47% (2636/5619) were neutral. We found that a mumps outbreak was observed right before November 2017, which may have encouraged people to discuss the importance of MMR vaccination. News articles highlighted this outbreak, for example, "Third dose of mumps vaccine could help stop outbreaks, researchers say" by PBS News Hour [71] and "CDC recommends booster shot of MMR vaccine during mumps outbreaks" by CNN [72] mentioned the outbreak and recommended the booster shot of MMR vaccine.

When tracking vaccine hesitancy, we found that the social media posts with a higher ratio of hesitancy were only observed in August 2014 (Figure 3C). During this month, some examples of articles could be associated with vaccine hesitancy: "Journal questions validity of autism and vaccine study" by CNN [73] and "Whistleblower Claims CDC Covered Up Data Showing Vaccine-Autism Link" by TIME [74]. While speculation, particularly among antivaccination subpopulations, continues to surround the discredited study linking MMR vaccines with autism, it is crucial to emphasize that this link has been unequivocally debunked by subsequent research, and organizations such as the CDC and WHO have clarified that no such association exists [75-77]. Nonetheless, these news articles, considered by some as antivaccine propaganda, may partially explain the observed trends in MMR vaccine hesitancy during August 2014.

Strengths and Limitations

In this study, we introduced an NLP-powered online monitoring tool for tracking vaccine-related discussions on multiple social media platforms, covering 3 vaccine topic groups. Our system provides several features that distinguish it from existing tools. It uses NLP algorithms to perform sentiment analysis on social media posts and facilitates the tracking of temporal trends and geographic clustering of vaccine sentiment and hesitancy through visualization. In addition, our system enables users to compare vaccine sentiment and hesitancy across different social media platforms. We have publicly shared our annotated social media vaccine corpora, and we have evaluated several text classification algorithms, providing a benchmark for future research. One of the hypothetical use cases is that our NLP-based tool's application spans from gauging vaccine sentiment during disease outbreaks to when a new vaccine is introduced. During an outbreak, the tool effectively analyzed sentiments toward measles vaccination, facilitating adjustments in public health campaigns.

While our proposed method uses the coarse-grained sentiment model (ie, represents the sentiment as a positive or negative class), fine-grained sentiment models, unlike traditional independent dimensional approaches, beneficially incorporate relations between dimensions, such as valence and arousal, into deep neural networks, thereby providing more nuanced, real-valued sentiment analysis and enhancing prediction accuracy [78-81]. These models prove particularly valuable in language-specific applications and are capable of classifying emotion categories and simultaneously predicting valence, arousal, and dominance scores for specific sentences, providing more nuanced sentiment analysis compared with simple positive or negative classifications.

Beyond the limitations inherent in the sentiment model, our approach also encounters constraints due to the use of traditional machine learning algorithms. Deep learning methods for word or sentiment embedding offer enhanced performance in sentiment analysis tasks by integrating external knowledge such as sentiment polarity and emotional semantics into word vectors [82-87]. They leverage neural networks and multitask learning to create task-specific embeddings, improving the accuracy of tasks such as sentiment and emotion analysis and sarcasm and stress detection [82-84,86]. Furthermore, these methods can adapt to the dynamic nature of language, handling out-of-vocabulary words and context-specific word meanings, proving more accurate and comprehensive than traditional word embeddings [86,87]. In future iterations, we plan to enrich our tool by integrating cutting-edge methods, alongside a more robust evaluation method such as time series cross-validation [88].

While previous studies have used NLP for sentiment analysis on COVID-19 vaccination and information exposure analysis regarding the HPV vaccine using Twitter data sets [40,89], and have investigated the temporal and geographic variations in public perceptions of the HPV vaccine [90], our tool extends its functionality to include a broader spectrum of platforms for tracking different vaccine sentiment and hesitancy on social media. Despite the scientific evidence supporting the safety and

efficacy of vaccines, vaccine hesitancy sentiments on social media can impact public confidence regarding vaccination [91]. Our tool is designed to quickly identify surges in vaccine hesitancy and thereby could be a tool to assist public health professionals in responding promptly with accurate information and effective vaccine promotion strategies.

However, it is essential to acknowledge the inherent limitations of using social media as a public health surveillance tool. These limitations include geography and language restrictions, as well as potential population, age, and gender biases, given that social media users may not represent the general population [92-94]. The user diversity across various social media platforms might partly account for the variation in sentiment and hesitancy label distributions. For example, YouTube has a high volume of users, but Twitter had the most activity in our study because people may view YouTube videos without leaving comments [93]. Moreover, owners of YouTube channels also have the option to disable comments on their uploaded videos. In addition, YouTube comments are highly tied to the content of the videos that the model might not have access to, leading to misinterpretations of sentiment and hesitancy. These biases and variabilities could partly account for the lower prediction accuracy observed for YouTube. Therefore, caution should be exercised when interpreting findings based on social media data, particularly considering the varying distributions of sentiment and hesitancy across different social media platforms in our study. Another limitation pertains to the absence of a weighting system in the dashboard. Currently, the impact of each post, considering variables such as the number of views or reposts, is not considered. In addition, private interactions, specifically on sites such as Facebook, might go unnoticed and this lack of access to private dialogues could limit the comprehensiveness of the responses we capture. Finally, there is the possibility of shifts in user behavior to emerging social media platforms, such as TikTok, introducing additional population bias if such platforms are not included in further analyses.

Conclusions

This study successfully developed an innovative real-time monitoring system for analyzing vaccine sentiment and hesitancy across 3 major social media platforms. This system uses NLP and machine learning to mine and classify social media discussions on vaccines, providing valuable insights into public sentiment and hesitancy trends. The application of this tool presents significant implications for public health strategies, aiding in promptly identifying and mitigating vaccine misinformation, enhancing vaccine uptake, and assisting in the execution of targeted health campaigns. Moreover, it encourages health care professionals to foster an evidence-based discourse around vaccines, thus counteracting misinformation and improving public health outcomes.

Acknowledgments

This work was funded by Merck Sharp & Dohme Corp, a subsidiary of Merck & Co, Inc (Rahway, New Jersey). The content is the sole responsibility of the authors and does not necessarily represent the official views of Merck & Co, Inc or Melax Tech.

Authors' Contributions

JD, ALE, and LY conceptualized and designed the study. LCH and JD performed the experiments. LCH, ALE, JD, and LY drafted the paper. LCH, LH, and JD performed the acquisition, analysis, or interpretation of data. All authors performed critical revision of the paper for important intellectual content. JD, ALE, and LY performed study supervision.

Conflicts of Interest

ALE is a current employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States, who may own stock and stock options in the Company. LY was an employee of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States during the time of the study. Melax Tech, including JW, JD, and LCH, was compensated for activities related to the execution of the study. FJM was employed by Melax Tech and IMO Health during the research described. IMO Health retains interests in certain software described in this article.

Multimedia Appendix 1
The online dashboard's user interface, architecture, and performances.
[DOCX File , 2065 KB - medinform_v12i1e57164_app1.docx]

References

1. Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. Lancet Infect Dis 2022;22(9):1293-1302 [FREE Full text] [doi: 10.1016/S1473-3099(22)00320-6] [Medline: 35753318]
2. Lindmeier C. Measles vaccination has saved an estimated 17.1 million lives since 2000. World Health Organization. 2015. URL: https://www.who.int/news/item/12-11-2015-measles-vaccination-has-saved-an-estimated-17-1-million-lives-since-2000 [accessed 2024-05-08]
3. Ehreth J. The global value of vaccination. Vaccine 2003;21(7-8):596-600. [doi: 10.1016/s0264-410x(02)00623-0] [Medline: 12531324]

4. MacDonald NE, SAGE Working Group on Vaccine Hesitancy. Vaccine hesitancy: definition, scope and determinants. Vaccine 2015;33(34):4161-4164 [FREE Full text] [doi: 10.1016/j.vaccine.2015.04.036] [Medline: 25896383]
5. Ten threats to global health in 2019. World Health Organization. URL: https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019 [accessed 2024-05-08]
6. Dubé E, Vivion M, MacDonald NE. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. Expert Rev Vaccines 2015;14(1):99-117. [doi: 10.1586/14760584.2015.964212] [Medline: 25373435]
7. Lane S, MacDonald NE, Marti M, Dumolard L. Vaccine hesitancy around the globe: analysis of three years of WHO/UNICEF joint reporting form data-2015-2017. Vaccine 2018;36(26):3861-3867 [FREE Full text] [doi: 10.1016/j.vaccine.2018.03.063] [Medline: 29605516]
8. Black FL. The role of herd immunity in control of measles. Yale J Biol Med 1982;55(3-4):351-360 [FREE Full text] [Medline: 7180027]
9. Cockman P, Dawson L, Mathur R, Hull S. Improving MMR vaccination rates: herd immunity is a realistic goal. BMJ 2011;343:d5703. [doi: 10.1136/bmj.d5703] [Medline: 21971162]
10. Lieu TA, Ray GT, Klein NP, Chung C, Kulldorff M. Geographic clusters in underimmunization and vaccine refusal. Pediatrics 2015;135(2):280-289. [doi: 10.1542/peds.2014-2715] [Medline: 25601971]
11. Omer SB, Pan WKY, Halsey NA, Stokley S, Moulton LH, Navar AM, et al. Nonmedical exemptions to school immunization requirements: secular trends and association of state policies with pertussis incidence. JAMA 2006;296(14):1757-1763 [FREE Full text] [doi: 10.1001/jama.296.14.1757] [Medline: 17032989]
12. Dempsey AF, Schaffer S, Singer D, Butchart A, Davis M, Freed GL. Alternative vaccination schedule preferences among parents of young children. Pediatrics 2011;128(5):848-856. [doi: 10.1542/peds.2011-0400] [Medline: 21969290]
13. Sadaf A, Richards JL, Glanz J, Salmon DA, Omer SB. A systematic review of interventions for reducing parental vaccine refusal and vaccine hesitancy. Vaccine 2013;31(40):4293-4304. [doi: 10.1016/j.vaccine.2013.07.013] [Medline: 23859839]
14. Zhao Z, Luman ET. Progress toward eliminating disparities in vaccination coverage among U.S. children, 2000-2008. Am J Prev Med 2010;38(2):127-137. [doi: 10.1016/j.amepre.2009.10.035] [Medline: 20117568]
15. Zimet GD, Weiss TW, Rosenthal SL, Good MB, Vichnin MD. Reasons for non-vaccination against HPV and future vaccination intentions among 19-26 year-old women. BMC Womens Health 2010;10:27 [FREE Full text] [doi: 10.1186/1472-6874-10-27] [Medline: 20809965]
16. Dredze M, Broniatowski DA, Smith MC, Hilyard KM. Understanding vaccine refusal: why we need social media now. Am J Prev Med 2016;50(4):550-552 [FREE Full text] [doi: 10.1016/j.amepre.2015.10.002] [Medline: 26655067]
17. Peretti-Watel P, Larson HJ, Ward JK, Schulz WS, Verger P. Vaccine hesitancy: clarifying a theoretical framework for an ambiguous notion. PLoS Curr 2015 Feb 25;7:ecurrents.outbreaks.6844c80ff9f5b273f34c91f71b7fc289 [FREE Full text] [doi: 10.1371/currents.outbreaks.6844c80ff9f5b273f34c91f71b7fc289] [Medline: 25789201]
18. Galagali PM, Kinikar AA, Kumar VS. Vaccine hesitancy: obstacles and challenges. Curr Pediatr Rep 2022;10(4):241-248 [FREE Full text] [doi: 10.1007/s40124-022-00278-9] [Medline: 36245801]
19. Larson HJ, Smith DMD, Paterson P, Cumming M, Eckersberger E, Freifeld CC, et al. Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. Lancet Infect Dis 2013;13(7):606-613. [doi: 10.1016/S1473-3099(13)70108-7] [Medline: 23676442]
20. Lawrence HY, Hausman BL, Dannenberg CJ. Reframing medicine's publics: the local as a public of vaccine refusal. J Med Humanit 2014;35(2):111-129. [doi: 10.1007/s10912-014-9278-4] [Medline: 24682632]
21. WHO T. The guide to tailoring immunization programmes. WHO Regional Office for Europe. 2013. URL: https://iris.who.int/handle/10665/351166 [accessed 2024-05-08]
22. Yaqub O, Castle-Clarke S, Sevdalis N, Chataway J. Attitudes to vaccination: a critical review. Soc Sci Med 2014;112:1-11 [FREE Full text] [doi: 10.1016/j.socscimed.2014.04.018] [Medline: 24788111]
23. Cox DS, Cox AD, Sturm L, Zimet G. Behavioral interventions to increase HPV vaccination acceptability among mothers of young girls. Health Psychol 2010;29(1):29-39. [doi: 10.1037/a0016942] [Medline: 20063933]
24. Cates JR, Ortiz R, Shafer A, Romocki LS, Coyne-Beasley T. Designing messages to motivate parents to get their preteenage sons vaccinated against human papillomavirus. Perspect Sex Reprod Health 2012;44(1):39-47 [FREE Full text] [doi: 10.1363/4403912] [Medline: 22405151]
25. Clayton EW, Hickson GB, Miller CS. Parents' responses to vaccine information pamphlets. Pediatrics 1994;93(3):369-372. [Medline: 8115193]
26. Kagashe I, Yan Z, Suheryani I. Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data. J Med Internet Res 2017;19(9):e315 [FREE Full text] [doi: 10.2196/jmir.7393] [Medline: 28899847]
27. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci 2015;26(2):159-169 [FREE Full text] [doi: 10.1177/0956797614557867] [Medline: 25605707]
28. Chan B, Lopez A, Sarkar U. The canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. PLoS One 2015;10(8):e0135072 [FREE Full text] [doi: 10.1371/journal.pone.0135072] [Medline: 26252774]

29. Mitra T, Counts S, Pennebaker J. Understanding anti-vaccination attitudes in social media. 2016 Presented at: Tenth International AAAI Conference on Web and Social Media; May 17-20, 2016; Cologne, Germany p. 269-278 URL: https://ojs.aaai.org/index.php/ICWSM/issue/view/272 [doi: 10.1609/icwsm.v10i1.14729]
30. McDonald L, Malcolm B, Ramagopalan S, Syrad H. Real-world data and the patient perspective: the promise of social media? BMC Med 2019;17(1):11 [FREE Full text] [doi: 10.1186/s12916-018-1247-8] [Medline: 30646913]
31. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
32. Becker BFH, Larson HJ, Bonhoeffer J, van Mulligen EM, Kors JA, Sturkenboom MCJM. Evaluation of a multinational, multilingual vaccine debate on Twitter. Vaccine 2016;34(50):6166-6171 [FREE Full text] [doi: 10.1016/j.vaccine.2016.11.007] [Medline: 27840012]
33. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL. The measles vaccination narrative in Twitter: a quantitative analysis. JMIR Public Health Surveill 2016;2(1):e1 [FREE Full text] [doi: 10.2196/publichealth.5059] [Medline: 27227144]
34. Love B, Himelboim I, Holton A, Stewart K. Twitter as a source of vaccination information: content drivers and what they are saying. Am J Infect Control 2013;41(6):568-570. [doi: 10.1016/j.ajic.2012.10.016] [Medline: 23726548]
35. Keelan J, Pavri V, Balakrishnan R, Wilson K. An analysis of the human papilloma virus vaccine debate on MySpace blogs. Vaccine 2010;28(6):1535-1540 [FREE Full text] [doi: 10.1016/j.vaccine.2009.11.060] [Medline: 20003922]
36. Chowdhary KR. Natural language processing. In: Fundamentals of Artificial Intelligence. New York City: Springer; 2020:603-649.
37. Vinet L, Zhedanov A. A 'missing' family of classical orthogonal polynomials. J Phys A Math Theor 2011;44(8):085201. [doi: 10.1088/1751-8113/44/8/085201]
38. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011;7(10):e1002199 [FREE Full text] [doi: 10.1371/journal.pcbi.1002199] [Medline: 22022249]
39. Vaccine Sentiments on Social Media. URL: https://vaccine.social/ [accessed 2024-05-08]
40. Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P. Covid-19 vaccine hesitancy: text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Syst Appl 2023;212:118715 [FREE Full text] [doi: 10.1016/j.eswa.2022.118715] [Medline: 36092862]
41. Kumar N, Corpus I, Hans M, Harle N, Yang N, McDonald C, et al. COVID-19 vaccine perceptions in the initial phases of US vaccine roll-out: an observational study on reddit. BMC Public Health 2022;22(1):446 [FREE Full text] [doi: 10.1186/s12889-022-12824-7] [Medline: 35255881]
42. Li HOY, Pastukhova E, Brandts-Longtin O, Tan MG, Kirchhof MG. YouTube as a source of misinformation on COVID-19 vaccination: a systematic analysis. BMJ Glob Health 2022;7(3):e008334 [FREE Full text] [doi: 10.1136/bmjgh-2021-008334] [Medline: 35264318]
43. Kwon S, Park A. Examining thematic and emotional differences across Twitter, Reddit, and YouTube: the case of COVID-19 vaccine side effects. Comput Human Behav 2023;144:107734 [FREE Full text] [doi: 10.1016/j.chb.2023.107734] [Medline: 36942128]
44. Aleksandric A, Anderson HI, Melcher S, Nilizadeh S, Wilson GM. Spanish Facebook posts as an indicator of COVID-19 vaccine hesitancy in Texas. Vaccines (Basel) 2022;10(10):1713 [FREE Full text] [doi: 10.3390/vaccines10101713] [Medline: 36298580]
45. Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. Knowl Inf Syst 2018;60(2):617-663. [doi: 10.1007/s10115-018-1236-4]
46. Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min 2021;11(1):81 [FREE Full text] [doi: 10.1007/s13278-021-00776-6] [Medline: 34484462]
47. Bonanni P, Bechini A, Donato R, Capei R, Sacco C, Levi M, et al. Human papilloma virus vaccination: impact and recommendations across the world. Ther Adv Vaccines 2015;3(1):3-12 [FREE Full text] [doi: 10.1177/2051013614557476] [Medline: 25553242]
48. Bankamp B, Hickman C, Icenogle JP, Rota PA. Successes and challenges for preventing measles, mumps and rubella by vaccination. Curr Opin Virol 2019;34:110-116. [doi: 10.1016/j.coviro.2019.01.002] [Medline: 30852425]
49. Eiden AL, DiFranzo A, Bhatti A, Wang HE, Bencina G, Yao L, et al. Changes in vaccine administration trends across the life-course during the COVID-19 pandemic in the United States: a claims database study. Expert Rev Vaccines 2023;22(1):481-494 [FREE Full text] [doi: 10.1080/14760584.2023.2217257] [Medline: 37218717]
50. Kleinbaum DG, Klein M, Pryor ER. Logistic Regression: A Self-Learning Text. Berlin, Heidelberg, Dordrecht, New York City: Springer; 2002.
51. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24(12):1565-1567. [doi: 10.1038/nbt1206-1565]
52. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens 2007;26(1):217-222. [doi: 10.1080/01431160412331269698]

53. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. New York, NY, United States: Association for Computing Machinery; 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, California, USA p. 785-794 URL: https://dl.acm.org/doi/proceedings/10.1145/2939672 [doi: 10.1145/2939672.2939785]
54. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. Proceedings VLDB Endowment 2017;11(3):269-282 [FREE Full text] [doi: 10.14778/3157794.3157797] [Medline: 29770249]
55. Schober P, Vetter TR. Logistic regression in medical research. Anesth Analg 2021;132(2):365-366 [FREE Full text] [doi: 10.1213/ANE.0000000000005247] [Medline: 33449558]
56. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng 2018;2018:4302425 [FREE Full text] [doi: 10.1155/2018/4302425] [Medline: 29849998]
57. Alam MZ, Rahman MS, Rahman MS. A random forest based predictor for medical data classification using feature ranking. Inform Med Unlocked 2019;15:100180 [FREE Full text] [doi: 10.1016/j.imu.2019.100180]
58. Alotaibi Y, Malik MN, Khan HH, Batool A, Alsufyani A, Alghamdi S, et al. Suggestion mining from opinionated text of big social media data. CMC-Comput Mater Con 2021;68(3):3323-3338 [FREE Full text] [doi: 10.32604/cmc.2021.016727]
59. Mallory EK, de Rochemonteix M, Ratner A, Acharya A, Re C, Bright RA, et al. Extracting chemical reactions from text using Snorkel. BMC Bioinformatics 2020;21(1):217 [FREE Full text] [doi: 10.1186/s12859-020-03542-1] [Medline: 32460703]
60. Ramos J. Using TF-IDF to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning. 2003. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c [accessed 2024-05-13]
61. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357 [FREE Full text] [doi: 10.1613/jair.953]
62. Tang Y, Zhang YQ, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern B Cybern 2009;39(1):281-288. [doi: 10.1109/TSMCB.2008.2002909] [Medline: 19068445]
63. Kirasich K, Smith T, Sadler B. Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review 2018;1(3):9.
64. Mehta H, Passi K. Social media hate speech detection using Explainable Artificial Intelligence (XAI). Algorithms 2022;15(8):291 [FREE Full text] [doi: 10.3390/a15080291]
65. Hensley S. Worried parents balk at HPV vaccine for daughters. NPR. 2013. URL: https://www.npr.org/sections/health-shots/2013/03/18/174617709/worried-parents-balk-at-hpv-vaccine-for-daughters [accessed 2024-05-08]
66. Castillo M. Side effect fears stop parents from getting HPV vaccine for daughters. CBS News. 2013. URL: https://www.cbsnews.com/news/side-effect-fears-stop-parents-from-getting-hpv-vaccine-for-daughters/ [accessed 2024-05-08]
67. Zimet GD, Rosberger Z, Fisher WA, Perez S, Stupiansky NW. Beliefs, behaviors and HPV vaccine: correcting the myths and the misinformation. Prev Med 2013;57(5):414-418 [FREE Full text] [doi: 10.1016/j.ypmed.2013.05.013] [Medline: 23732252]
68. Karafillakis E, Simas C, Jarrett C, Verger P, Peretti-Watel P, Dib F, et al. HPV vaccination in a context of public mistrust and uncertainty: a systematic literature review of determinants of HPV vaccine hesitancy in Europe. Hum Vaccin Immunother 2019;15(7-8):1615-1627 [FREE Full text] [doi: 10.1080/21645515.2018.1564436] [Medline: 30633623]
69. HPV, the vaccine for HPV, and cancers caused by HPV. Centers for Disease Control and Prevention. 2022. URL: https://tinyurl.com/2d45j3jz [accessed 2024-05-08]
70. Meites E, Szilagyi PG, Chesson HW, Unger ER, Romero JR, Markowitz LE. Human papillomavirus vaccination for adults: updated recommendations of the advisory committee on immunization practices. MMWR Morb Mortal Wkly Rep 2019;68(32):698-702 [FREE Full text] [doi: 10.15585/mmwr.mm6832a3] [Medline: 31415491]
71. Branswell H. Third dose of mumps vaccine could help stop outbreaks, researchers say. STAT. 2017. URL: https://www.statnews.com/2017/09/06/mumps-vaccine-study/ [accessed 2024-05-08]
72. Scutti S. CDC recommends booster shot of MMR vaccine during mumps outbreaks. CNN Health. 2017. URL: https://www.cnn.com/2017/10/25/health/cdc-mumps-outbreak-syracuse-university/index.html [accessed 2024-05-08]
73. Goldschmidt D. Journal questions validity of autism and vaccine study. CNN Health. 2014. URL: https://www.cnn.com/2014/08/27/health/irpt-cdc-autism-vaccine-study/index.html [accessed 2024-05-08]
74. Park A. Whistleblower claims CDC covered up data showing vaccine-autism link. TIME. 2014. URL: https://time.com/3208886/whistleblower-claims-cdc-covered-up-data-showing-vaccine-autism-link/ [accessed 2024-05-08]
75. Autism and vaccines. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/vaccinesafety/concerns/autism.html [accessed 2024-05-08]
76. Epidemiological WW. MMR and autism. World Health Organization. 2003. URL: https://www.who.int/groups/global-advisory-committee-on-vaccine-safety/topics/mmr-vaccines-and-autism [accessed 2024-05-08]
77. The Editors of The Lancet, Caplan AL. Retraction—ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 2010;375(9713):445. [doi: 10.1016/s0140-6736(10)60175-4]

78. Xie H, Lin W, Lin S, Wang J, Yu LC. A multi-dimensional relation model for dimensional sentiment analysis. Inf Sci 2021;579:832-844 [FREE Full text] [doi: 10.1016/j.ins.2021.08.052]
79. Park S, Kim J, Ye S, Jeon J, Park HY, Oh A. Dimensional emotion detection from categorical emotion. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; Nov 2021:4367-4380.
80. Lee LH, Li JH, Yu LC. Chinese EmoBank: building valence-arousal resources for dimensional sentiment analysis. ACM Trans Asian Low-Resour Lang Inf Process 2022;21(4):1-18 [FREE Full text] [doi: 10.1145/3489141]
81. Wang J, Yu LC, Lai KR, Zhang X. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. IEEE/ACM Trans Audio Speech Lang Process 2020;28:581-591 [FREE Full text] [doi: 10.1109/taslp.2019.2959251]
82. Tang D, Wei F, Qin B, Yang N, Liu T, Zhou M. Sentiment embeddings with applications to sentiment analysis. IEEE Trans Knowl Data Eng 2016;28(2):496-509. [doi: 10.1109/tkde.2015.2489653]
83. Xu P, Madotto A, Wu CS, Park JH, Fung P. Emo2Vec: learning generalized emotion representation by multi-task training. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels, Belgium: Association for Computational Linguistics; Oct 2018:292-298.
84. Ye Z, Li F, Baldwin T. Encoding sentiment information into word vectors for sentiment analysis. : Association for Computational Linguistics; 2018 Presented at: Proceedings of the 27th International Conference on Computational Linguistics; August, 2018; Santa Fe, New Mexico, USA URL: https://aclanthology.org/C18-1085/
85. Yu LC, Wang J, Lai KR, Zhang X. Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans Audio Speech Lang Process 2018;26(3):671-681. [doi: 10.1109/taslp.2017.2788182]
86. Wang J, Zhang Y, Yu LC, Zhang X. Contextual sentiment embeddings via bi-directional GRU language model. Knowl-Based Syst 2022;235:107663 [FREE Full text] [doi: 10.1016/j.knosys.2021.107663]
87. Zhu L, Li W, Shi Y, Guo K. SentiVec: learning sentiment-context vector via kernel optimization function for sentiment analysis. IEEE Trans Neural Netw Learn Syst 2021;32(6):2561-2572. [doi: 10.1109/TNNLS.2020.3006531] [Medline: 32673198]
88. Bergmeir C, Hyndman RJ, Koo B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput Stat Data An 2018;120:70-83. [doi: 10.1016/j.csda.2017.11.003]
89. Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. Vaccine 2017;35(23):3033-3040 [FREE Full text] [doi: 10.1016/j.vaccine.2017.04.060] [Medline: 28461067]
90. Du J, Luo C, Shegog R, Bian J, Cunningham RM, Boom JA, et al. Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. JAMA Netw Open 2020;3(11):e2022025 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.22025] [Medline: 33185676]
91. Zhang Q, Zhang R, Wu W, Liu Y, Zhou Y. Impact of social media news on COVID-19 vaccine hesitancy and vaccination behavior. Telemat Inform 2023;80:101983 [FREE Full text] [doi: 10.1016/j.tele.2023.101983] [Medline: 37122766]
92. Zhao Y, He X, Feng Z, Bost S, Prosperi M, Wu Y, et al. Biases in using social media data for public health surveillance: a scoping review. Int J Med Inform 2022;164:104804. [doi: 10.1016/j.ijmedinf.2022.104804] [Medline: 35644051]
93. Auxier B, Anderson M. Social media use in 2021. Pew Research Center 2021;1:1-4 [FREE Full text] [doi: 10.4135/9781412963947.n376]
94. Shor E, van de Rijt A, Fotouhi B. A large-scale test of gender bias in the media. SocScience 2019;6:526-550 [FREE Full text] [doi: 10.15195/v6.a20]

Abbreviations

3Cs: confidence, complacency, and convenience
API: application programming interface
CDC: Centers for Disease Control and Prevention
HPV: human papillomavirus
LR: logistic regression
MMR: measles, mumps, and rubella
NLP: natural language processing
SMOTE: synthetic minority over-sampling technique
SVM: support vector machine
VPD: vaccine-preventable disease
WHO: World Health Organization
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 07.02.24; peer-reviewed by M Chatzimina, S Lee, LC Yu, X Vargas Meza; comments to author 25.03.24; revised version received 08.04.24; accepted 11.04.24; published 21.06.24.

Please cite as:
Huang LC, Eiden AL, He L, Annan A, Wang S, Wang J, Manion FJ, Wang X, Du J, Yao L
Natural Language Processing–Powered Real-Time Monitoring Solution for Vaccine Sentiments and Hesitancy on Social Media: System Development and Validation
JMIR Med Inform 2024;12:e57164
URL: https://medinform.jmir.org/2024/1/e57164
doi:10.2196/57164
PMID:38904984

Original Paper

Data-Driven Identification of Factors That Influence the Quality of Adverse Event Reports: 15-Year Interpretable Machine Learning and Time-Series Analyses of VigiBase and QUEST

Sim Mei Choo[1,2], MSc; Daniele Sartori[3], MSc; Sing Chet Lee[1], MSc; Hsuan-Chia Yang[2,4,5,6*], PhD; Shabbir Syed-Abdul[2,4,7*], MD, PhD

[1]Centre of Compliance & Quality Control, National Pharmaceutical Regulatory Agency, Petaling Jaya, Malaysia

[2]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan

[3]Uppsala Monitoring Centre, Uppsala, Sweden

[4]International Center for Health Information Technology, Taipei Medical University, Taipei, Taiwan

[5]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan

[6]Research Center of Big Data and Meta-Analysis, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

[7]School of Gerontology and Long-Term Care, Taipei Medical University, Taipei, Taiwan

[*]these authors contributed equally

Corresponding Author:
Shabbir Syed-Abdul, MD, PhD
Graduate Institute of Biomedical Informatics
Taipei Medical University
301 Yuantong Rd
Taipei, 235
Taiwan
Phone: 886 66202589 ext 10930
Email: drshabbir@tmu.edu.tw

Abstract

Background: The completeness of adverse event (AE) reports, crucial for assessing putative causal relationships, is measured using the vigiGrade completeness score in VigiBase, the World Health Organization global database of reported potential AEs. Malaysian reports have surpassed the global average score (approximately 0.44), achieving a 5-year average of 0.79 (SD 0.23) as of 2019 and approaching the benchmark for well-documented reports (0.80). However, the contributing factors to this relatively high report completeness score remain unexplored.

Objective: This study aims to explore the main drivers influencing the completeness of Malaysian AE reports in VigiBase over a 15-year period using vigiGrade. A secondary objective was to understand the strategic measures taken by the Malaysian authorities leading to enhanced report completeness across different time frames.

Methods: We analyzed 132,738 Malaysian reports (2005-2019) recorded in VigiBase up to February 2021 split into historical International Drug Information System (INTDIS; n=63,943, 48.17% in 2005-2016) and newer E2B (n=68,795, 51.83% in 2015-2019) format subsets. For machine learning analyses, we performed a 2-stage feature selection followed by a random forest classifier to identify the top features predicting well-documented reports. We subsequently applied tree Shapley additive explanations to examine the magnitude, prevalence, and direction of feature effects. In addition, we conducted time-series analyses to evaluate chronological trends and potential influences of key interventions on reporting quality.

Results: Among the analyzed reports, 42.84% (56,877/132,738) were well documented, with an increase of 65.37% (53,929/82,497) since 2015. Over two-thirds (46,186/68,795, 67.14%) of the Malaysian E2B reports were well documented compared to INTDIS reports at 16.72% (10,691/63,943). For INTDIS reports, higher pharmacovigilance center staffing was the primary feature positively associated with being well documented. In recent E2B reports, the top positive features included reaction abated upon drug dechallenge, reaction onset or drug use duration of <1 week, dosing interval of <1 day, reports from public specialist hospitals, reports by pharmacists, and reaction duration between 1 and 6 days. In contrast, reports from product registration holders and other health care professionals and reactions involving product substitution issues negatively affected the quality of E2B reports. Multifaceted strategies and interventions comprising policy changes, continuity of education, and

human resource development laid the groundwork for AE reporting in Malaysia, whereas advancements in technological infrastructure, pharmacovigilance databases, and reporting tools concurred with increases in both the quantity and quality of AE reports.

Conclusions: Through interpretable machine learning and time-series analyses, this study identified key features that positively or negatively influence the completeness of Malaysian AE reports and unveiled how Malaysia has developed its pharmacovigilance capacity via multifaceted strategies and interventions. These findings will guide future work in enhancing pharmacovigilance and public health.

(JMIR Med Inform 2024;12:e49643) doi:10.2196/49643

KEYWORDS

pharmacovigilance; medication safety; big data analysis; feature selection; interpretable machine learning

Introduction

Background

Pharmacovigilance (PV) is the science and activities related to the detection, assessment, understanding, and prevention of adverse effects or any other possible drug-related problems [1]. Individual case safety reports (ICSRs) of suspected adverse drug reactions and adverse events following immunization (hereafter collectively referred to as adverse events [AEs]) collected in spontaneous reporting systems (SRSs) remain the cornerstone of postmarketing drug safety surveillance [2,3] (see Multimedia Appendix 1 for a list of definitions [4-10]).

Over 170 participating countries in the World Health Organization (WHO) Programme for International Drug Monitoring (PIDM) share reports of suspected AEs and collaborate worldwide in monitoring and identifying signals of AEs [4]. The WHO PIDM signal detection process is anchored on data recorded in the WHO global ICSR database, VigiBase, developed and maintained by the WHO Collaborating Centre for International Drug Monitoring, Uppsala Monitoring Centre (UMC), Sweden. Common technical specifications for report transmission and standard terminologies for drugs and reactions have evolved over the years to facilitate global information sharing and efficient analysis [4,5]. Currently, VigiBase accepts 3 standard formats: the original International Drug Information System (INTDIS) and 2 revisions of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) Guidelines for the Electronic Transmission of ICSRs, namely E2B(R2), and the latest E2B(R3), with all data being transformed to a format most closely resembling E2B(R2) in VigiBase [4].

The participating members are characterized by diverse contexts—sociocultural, political, and clinical—that affect the measures in which reports are collected and processed, as well as their quality [4]. Ideally, a robust PV system should consider all data quality parameters, including accuracy, completeness, conformity, consistency, currency, duplication, integrity, precision, relevance, and understandability. Among all these parameters, problems associated with completeness (ie, missing data) have long been regarded as critical factors hampering the usefulness of existing reports [5,6]. Influxes of poorly documented reports could increase operational burdens, upsurge a system's resources, and even mask or delay the detection of drug safety signals [11].

While only 4 elements (identifiable patient, identifiable reporter, medicinal product, and AE) are required for a valid report, they are often insufficient for productive analyses of potential causal relationships between medicinal products and AEs [6]. In 2014, the UMC developed the vigiGrade completeness score, an automated multidimensional tool that measures the amount of clinically relevant information in reports essential for causality assessment, replacing the 4-grade WHO documentation grading scheme since the 1990s [6,12]. The vigiGrade score quantifies report completeness based on a selection of ICH-E2B fields: time to onset, indication, event outcome, patient age and sex, dose information, country of origin, reporter, type of report, and free-text fields. The vigiGrade score can be used to pinpoint trends in report quality over time and reflect systematic data quality issues in collections of reports from member countries. For instance, vigiGrade uncovered miscoded age units in US reports and missing AE outcomes in Italian reports [6]. The score may also guide reviewers in judging whether the information in a report suffices for a problem to be investigated [4]. Notably, vigiGrade has proven to be an indicator of a true signal and is part of the data-driven predictive model used by the UMC, vigiRank, for signal detection [13].

The PV System in Malaysia

PV activities in Malaysia began in the 1980s with the establishment of the Malaysian Adverse Drug Reactions Advisory Committee (MADRAC) under the Drug Control Authority (DCA) [7]. The Malaysian national PV center is based within the National Pharmaceutical Regulatory Agency (NPRA) under the Pharmaceutical Services Programme of the Ministry of Health (MOH). Malaysia became a member of the WHO PIDM in 1990 and is regarded as an established PV center, receiving >30,000 reports annually, which is well above the WHO criteria of 200 reports per million inhabitants per year since 2009. Every AE report recorded in the national PV database (QUEST; see Multimedia Appendix 1 [7] for a detailed description) is carefully processed and assessed by trained pharmacists at the national center and subsequently reviewed by the MADRAC before submission to the UMC for inclusion in VigiBase (Figure S1 in Multimedia Appendix 2).

Problem Statement and Research Benefits

Previous studies have not evaluated the quality of Malaysian AE reports, and little is known about the underlying factors affecting their vigiGrade completeness scores. However, identifying and validating factors associated with report quality

was made difficult by the large number and variety of potentially correlated characteristics of a spontaneous report—at the reaction, drug, patient, reporter, sender, or regulator level (see the literature review on AE report quality in Multimedia Appendix 3 [6,14-38]). As of 2019, Malaysian reports demonstrated a 5-year average completeness score of 0.79, surpassing the global average of approximately 0.44 in VigiBase and approaching the benchmark for well-documented reports (0.80). Therefore, this study primarily aimed to use a hypothesis-free, data-driven approach to explore the main drivers influencing the completeness of Malaysian reports in VigiBase over a 15-year period using vigiGrade. A secondary objective was to understand the strategic measures taken by the Malaysian authorities that preceded the relatively high completeness score across different time frames. A better understanding of the drivers of AE report completeness may be helpful for the NPRA and regulators worldwide.

Methods

Data-Driven Framework for Identifying Factors Associated With AE Report Quality

Overview

Our study used big data analysis approaches incorporating machine learning (ML) methods, which are becoming increasingly prevalent in clinical and epidemiological research [8,39]. These approaches aimed to overcome limitations inherent in traditional approaches in handling complex interactions among variables (eg, multicollinearity and nonlinearity). Importantly, ML methods focus on identifying patterns and associations within complex data rather than on establishing causal inference [39]. For clarity, we have outlined the similarities in concepts and nomenclatures between ML and traditional medical statistics in Multimedia Appendix 1.

Hybrid Feature Selection

Our study leveraged ML methods for feature selection, mitigating human bias in analyzing extensive report characteristics, which might be overlooked by traditional hypothesis-driven approaches that are prone to high selection biases [9]. We combined statistical filtering and ML algorithms to preselect features, reducing overfitting risks (often arising from redundancy and multicollinearity) and computational costs [40,41]. Notably, Stevens et al [9] used random forest (RF)–based feature selection to identify potential risk factors associated with cardiovascular diseases. In almost all domains, incorporating domain expertise remains vital for developing meaningful and effective models [8]. Multimedia Appendix 4 [8,9,39-48] provides detailed explanations.

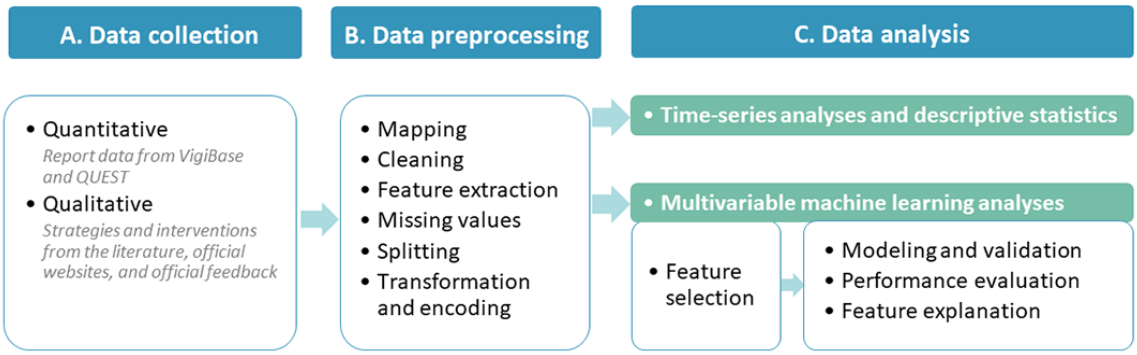
Interpretable ML

Post hoc explanation methods such as Shapley additive explanations (SHAP) have been increasingly used to provide interpretability for complex black-box models such as RF [49]. Van den Bosch et al [50] used regression coefficients and SHAP value analyses to identify risk factors associated with 30-day mortality among patients undergoing colorectal cancer surgery. Gong et al [51] also developed an ML framework for acute kidney injury prediction and interpretation using SHAP values to assess feature contributions and identify specific patient impacts. Multimedia Appendix 5 [49-57] provides detailed explanations.

Study Design

This observational study used interpretable ML and descriptive time-series analyses of PV database data. Fundamentally, it was an exploratory data analysis assessing a large number of report characteristics without prespecified hypotheses [42] aiming to identify factors influencing report quality. The main steps of our methodological workflow are illustrated in Figure 1.

Figure 1. Overview of study workflow.



Data Collection

AE report data were obtained from VigiBase in CSV format. Supplementary information, such as means of reporting, sender type, and sender region, was retrieved from the NPRA QUEST3+ (latest version) database in CSV format. For the secondary objective of understanding the key drivers of AE report completeness, strategies and interventions implemented in Malaysia were collected from the literature, official websites, and feedback from the NPRA. We included all reports recorded in VigiBase as of February 2021; reported in Malaysia; and received by the NPRA from January 1, 2005, to December 31, 2019. This 15-year range was chosen for its relevance to current PV needs and to enable the timely identification of necessary improvements. We excluded reports that (1) were suspected duplicates identified by the UMC's vigiMatch [4] (see Multimedia Appendix 1 for operational definitions), (2) were not sourced from Malaysia, (3) had null average completeness scores, and (4) lacked drugs marked as suspected or interacting.

Study Variables

Dependent Variables or Outcomes

The vigiGrade completeness score (*C*; ranges from 0.07 to 1) was classified as well documented ($C>0.8$) or not well documented ($C\leq0.8$; see Multimedia Appendix 1 for operational definitions).

Independent Variables or Explanatory Features

The variables related to administrative, sender, reporter, patient, drug, and reaction characteristics are presented in Table S1 in Multimedia Appendix 2.

Data Preprocessing

Data Mapping

Supplementary data from QUEST3+ were mapped to the primary data set from VigiBase using the primary identifier. Given the distinct differences in reporting elements of INTDIS and E2B formats (input values vary in certain data fields), we divided the data set for separate analysis.

Data Cleaning and Feature Extraction

We cleaned and engineered the features from the available data based on the literature, domain knowledge, and previous experience. Information about the WHO Anatomical Therapeutic Chemical (ATC) classification system codes was provided by the UMC in the data set. If an active ingredient was linked to more than one code, an ATC level-2 code was manually assigned based on indication, route of administration, dosage, product information, and clinical narratives. Reporting qualifications in INTDIS format were harmonized with the E2B format with reference to supplementary data from the NPRA. We calculated the number of suspected or interacting drugs, concomitant drugs, and reactions for each report. We also included the annual staffing level of the national PV center in Malaysia and the means of reporting (based on report identifier).

Missing Values

Continuous variables consisting of null values were converted to categorical variables based on data distribution and domain knowledge. Missing values for categorical variables were grouped as a *null* category.

Data Splitting

To ensure a consistent distribution of target classes, we applied stratified random sampling. We allocated 90% of the data for training, which underwent 10-fold cross-validation, and reserved 10% for testing to gauge model performance on unseen samples. Our approach prioritized the extraction of insights from the current data set rather than overgeneralizations on future data.

Transformation and Encoding

To overcome data complexity and maximize interpretability, data at the drug event level were transformed to the case (report) level. Observations related to concomitant drugs were excluded as the vigiGrade scoring method is restricted to drugs listed as suspected or interacting [6]. We took the average value of a case for continuous variables whereby, for categorical variables, we examined the presence (or absence) of a particular drug- or event-related characteristic. Continuous variables were standardized. Binary categorical variables such as patient sex were integer encoded. One-hot encoding was performed on the remaining categorical variables, including ordinal variables and categories labelled as *null* or *unknown*. In the following sections, we distinguish variables from features, where the latter correspond to the processed variables in a binary fashion for the ML model input [43,56].

Multivariable ML Analysis

Feature Selection

We performed hybrid feature selection to eliminate redundant or less informative features before data mining using the ML algorithm. To avoid data leakage and the corresponding model overfitting, we conducted a 2-stage feature selection solely based on training data [8,44]. We first applied the univariable filter method to independently assess and preselect the features and subsequently selected the top-ranked features using RF-based recursive feature elimination coupled with multicollinearity assessment. The detailed processes are provided in Multimedia Appendix 4.

Modeling and Validation

We applied a supervised ML method to identify key features relevant to the reports classified as well documented. Specifically, the RF classifier was selected for its robustness to nonparametric distributions, nonlinearity, and outliers [58] and its out-of-the-box performance. Its built-in feature importance metrics allowed us to assess the relative attribution of a feature to the classification task. The more a feature is used to make key decisions with the forest of decision trees, the higher its relative importance. To mitigate class imbalance in the INTDIS data set, we used RandomUnderSampler with a 0.25 ratio that achieved optimal balanced performance of prediction and recall. We chose undersampling over synthetic sampling methods to preserve the real-world data characteristics. For the imbalanced INTDIS data set, we adjusted the class_weight parameter in the RF classifier to *balanced*. We evaluated the RF classification models using 10-fold cross-validation.

Performance Evaluations

Classification performance was measured using the area under the receiver operating characteristic curve, accuracy, recall (sensitivity), and precision (positive predictive values). For the imbalanced INTDIS data set, F_1-scores (harmonic average of precision and recall) were reported.

Feature Explanations

To mitigate the issue of black-box predictions**,** we used TreeExplainer [52] to generate SHAP summary plots that succinctly display the magnitude, prevalence, and direction of a feature's effect by measuring each feature's attributions to the classification. In SHAP, the feature effect is a measure of how much the value of a specific feature influences the prediction made by the model.

Software and Packages

All ML analyses were developed in Jupyter Notebook (Project Jupyter) using Python (version 3.7.9; Python Software

Foundation). Statistical tests were performed using *pandas* (version 1.2.4) and *statsmodels* (version 0.12.2) [59]. ML analysis was completed using the *scikit-learn* package (version 0.24.2) [60]. SHAP values were calculated using TreeExplainer [52].

Time-Series and Descriptive Statistical Analysis

We used time-series analysis and descriptive statistics to evaluate the trends in report quality and the characteristics associated with well-documented reports over different time frames. One-way ANOVA or 2-tailed Student *t* tests were conducted on continuous variables, whereas the chi-square or Fisher exact test was used to compare categorical variables, as appropriate. A *P* value of <.05 was considered statistically significant. All analyses were conducted using the SAS software (version 9.4; SAS Institute).

Ethical Considerations

This study was registered with the Malaysian National Medical Research Register (NMRR-20-983-53984 [Investigator Initiated Research]) and received ethics approval from the Medical Review and Ethics Committee, MOH, Malaysia (reference: KKM/NIHSEC/P20–1144(4)).

Results

Overview

We analyzed the completeness of Malaysian AE reports in VigiBase received by the NPRA over 15 years. A total of 132,738 reports were included in the analysis following the predefined inclusion and exclusion criteria (Figure S2 in Multimedia Appendix 2). Table S1 in Multimedia Appendix 2 summarizes the characteristics of the INTDIS and E2B reports included in this study concerning administration, reporter, patient, drug, and reaction by status of being well documented. Among the included reports, 48.17% (63,943/132,738) were in the INTDIS format, and 51.83% (68,795/132,738) were in the E2B format. Over two-thirds (46,186/68,795, 67.14%) of E2B reports were well documented compared to 16.72% (10,691/63,943) of INTDIS reports.

Multivariable ML Analysis

Selected Features

For the INTDIS subsets, 90 features were preselected using univariate filter methods and further narrowed down to 33 features following RF-based recursive feature elimination ranking and multicollinearity assessment. For the E2B subsets, 90 features were preselected and subsequently reduced to 40.

Classification Performance

The performance of the RF models in classifying reports as well or not well documented is presented in Table 1.

Table 1. Classification performance of random forest model for the training (10-fold cross-validation) and test set.

	Recall (%)	Precision (%)	Accuracy (%)	AUROC[a] (%)	F_1-score (%)
INTDIS[b]					
Training, mean (SD)	99.6 (0.03)	95.4 (0.13)	99.0 (0.03)	99.8 (0.01)	97.4 (0.06)
Validation, mean (SD)	73.7 (1.90)	77.0 (1.02)	90.3 (0.39)	95.0 (0.33)	75.3 (1.16)
Test	74.9	74.3	91.5	95.1	74.6
E2B					
Training, mean (SD)	99.7 (0.02)	99.7 (0.02)	99.6 (0.01)	99.9 (0.001)	99.7 (0.01)
Validation, mean (SD)	96.9 (0.27)	91.6 (0.34)	92.0 (0.24)	94.9 (0.29)	94.2 (0.20)
Test	96.9	90.9	91.4	95.1	93.8

[a]AUROC: area under the receiver operating characteristic curve.

[b]INTDIS: International Drug Information System.

Top Factors Predicting Status of Malaysian Reports Being Well Documented

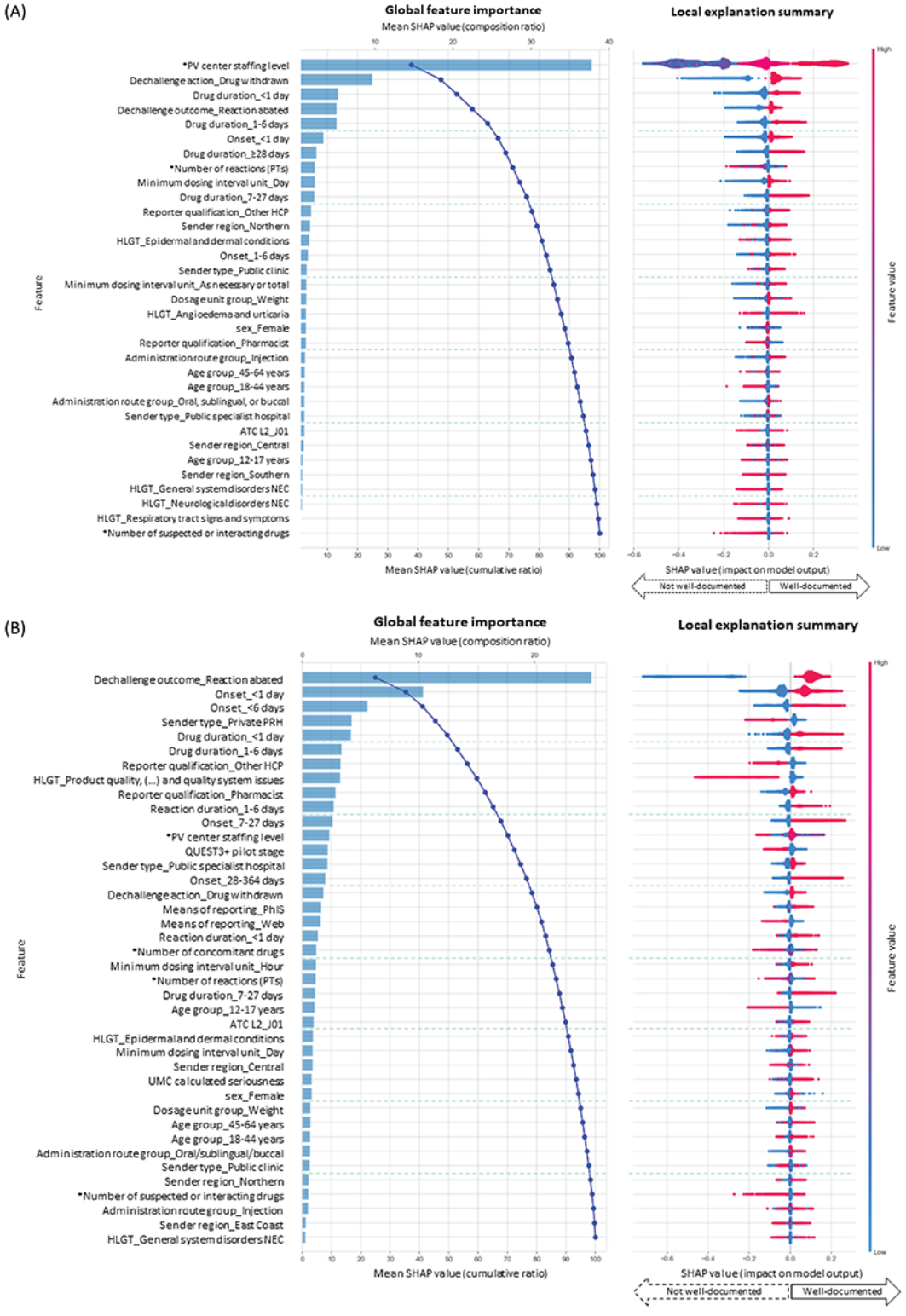
RF ML Model

Figure S3 in Multimedia Appendix 2 reveals the top-ranked features that contributed to the status of INTDIS and E2B reports being well documented derived from the RF ML model's built-in feature importance metrics. However, the directions of their contribution were not known due to the black-box nature of the RF model. For INTDIS reports received between 2005 and 2016, PV center staffing was identified as the most important factor in predicting their status of being well documented. Other important factors included suspected drug withdrawal, the number of reactions reported, reaction abated upon drug dechallenge, and patient sex. Reaction abated upon drug dechallenge, on the other hand, appeared to be the most important factor predicting whether an E2B report received between 2015 and 2019 was well documented. Reports from other health care professionals (HCPs), reactions occurring <1 day, reports submitted by product registration holders (PRHs), and the number of concomitant drugs were also among the top 5 important factors.

SHAP Interpretation Method

The SHAP post hoc interpretations provided us with an understanding of the magnitude, prevalence, and direction of feature effects. Figure 2 depicts rich summaries of individual attributions for all features, allowing us to discover key factors that influence well-documented reports. Features with higher global importance have a greater influence on the model's predictions. A feature with predominantly red dots to the right (eg, reaction abated upon drug dechallenge in Figure 2B) implies a positive contribution to well-documented reports, whereas a negative contribution is indicated if the direction is to the left (eg, reports submitted by PRHs in Figure 2B). Features with lower global importance but a long tail stretching in one direction indicate a rare but high-magnitude effect [52,56]. As mean SHAP values are calculated across all cases, a feature with a lower impact but higher prevalence may have a higher SHAP value [50]. For the least globally important features, we observed that their feature effects were not constant across cases, with blue and red dots dispersed in both directions. This variation may arise from interactions with other features that modulate their importance in different cases [52].

Figure 2. (A) Top 33 features for the International Drug Information System (INTDIS) subset from 2005 to 2016; (B) top 40 features for the E2B subset during the years 2015 to 2019. The Shapley additive explanation (SHAP) bar plot illustrates global feature importances based on mean absolute SHAP values, highlighting the impact of each feature on the model's predictions. Higher values represent greater influence. The waterfall plot indicates the cumulative contribution of features to the model. The SHAP summary plot of local explanations displays each observation as a dot, with its position on the x-axis (SHAP value) indicating the impact of a feature on the model's classification for that observation. Continuous features (marked with an asterisk) range from low (blue) to high (red) values, whereas categorical features of a binary nature are either absent (blue) or present (red). The distribution of dots indicates the magnitude and prevalence of a feature effect. Features are ordered by global importance. ATC: Anatomical Therapeutic Chemical; HCP: health care professional; HLGT: Medical Dictionary for Regulatory Activities High-Level Group Term; NEC: not elsewhere classified; PhIS: pharmacy hospital information system; PRH: product registration holder; PT: Medical Dictionary for Regulatory Activities Preferred Term; PV: pharmacovigilance; UMC: Uppsala Monitoring Centre; UMC calculated seriousness: serious cases classified automatically by a UMC-developed algorithm.



Regarding INTDIS reports, in earlier years, PV center staffing was the primary factor driving the Malaysian rate of reports being well documented, with this factor alone accounting for >35% of the model's explainability. The next most important factor favoring well-documented reports was drug withdrawal, followed by a duration of drug use of <1 week, reaction abated upon drug dechallenge, and reaction occurring <1 day. In contrast, an increased number of reported reactions and reports from pharmacists predicted not well-documented INTDIS reports.

In more recent years, the most important factor favoring well-documented E2B reports from Malaysia was reaction abated upon drug dechallenge, which alone was responsible for >25% of the model's explainability. Among the top 25 features, which provided 90% of the model's interpretation on classifying status of being well documented, 6 (24%) were found to be negatively associated with well-documented reports: reports submitted by PRHs; reports made by other HCPs; reactions under the Medical Dictionary for Regulatory Activities (MedDRA) High-Level Group Terms (HLGTs) *product quality, supply, distribution, manufacturing,* and *quality system issues*; reports received during the QUEST3+ pilot stage, reports received via web reporting, and adolescent patients (aged 12-17 years). E2B reports that involved reactions with a shorter time to onset and duration of drug use were more likely to be well documented. Other identified key drivers of Malaysian well-documented E2B reports were reports made by pharmacists, reports submitted from public specialist hospitals, pharmacy hospital information system (PhIS)–integrated reporting, and the involvement of systemic antimicrobials (ATC code J01).

Time-Series and Descriptive Statistical Analysis

Trend Analysis of Malaysian AE Report Quality (2005-2019)

Figure 3 depicts the time trends in AE reporting in Malaysia, illustrating how both the quantity and completeness scores of AE reports received by the NPRA grew over a 15-year period from 2005 to 2019. In Tables 2 and 3, we summarize the trends divided into 5-year subperiods, further stratified by sender type and reporter qualification. Of the total 132,738 reports received, 56,877 (42.84%) were well documented. Before 2014, the average completeness score consistently fell short of 0.5 but was slightly above the global average of 0.44. The volume of reports surged by 121% from 2013 to 2014, whereas the proportion of well-documented reports rose from practically 0% to 18.93% (2843/15,013). Since 2015, more than half (53,929/82,497, 65.37%) of the Malaysian reports were well documented, averaging 0.79 (SD 0.23) over the last 5 years, with a new high of 0.82 in 2019.

Figure 3. Distribution of average completeness scores and counts of Malaysian reports by status of being well documented in VigiBase over the study period.

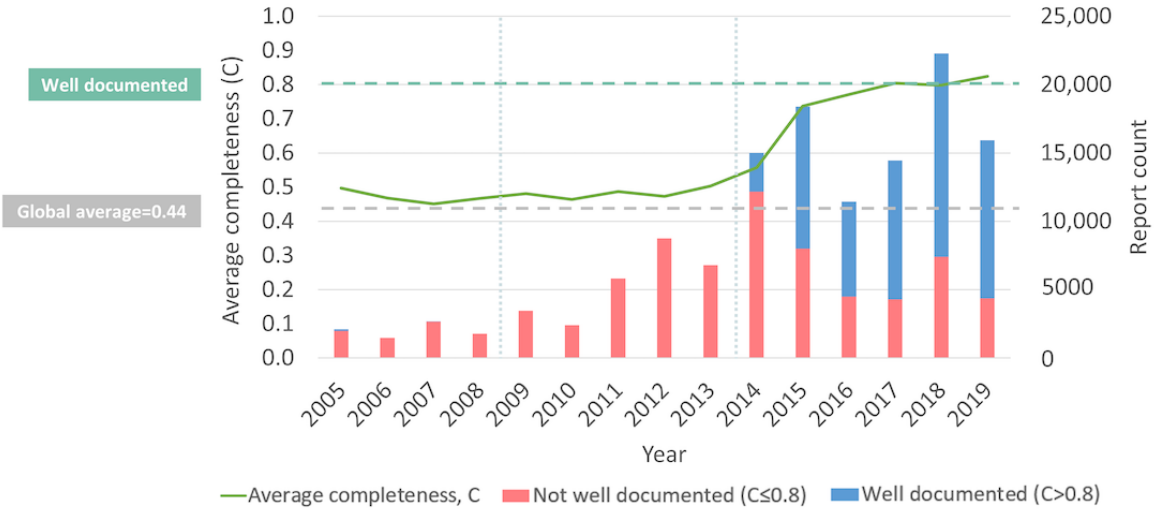


Table 2. The 5-year stratified summary statistics of overall Malaysian report quality.

	Total	2005-2009	2010-2014	2015-2019	*P* value[a]
Overall reports, n (%)	132,738 (100)	11,458 (8.6)	38,783 (29.2)	82,497 (62.2)	N/A[b]
Completeness, mean (SD)	0.68 (0.25)	0.47 (0.15)	0.51 (0.18)	0.79 (0.23)	<.001
Well-documented reports (C^c>0.8), n (%)	56,877 (42.8)	105 (0.9)	2843 (7.3)	53,929 (65.4)	<.001

[a]*P* value based on ANOVA.

[b]N/A: not applicable.

[c]vigiGrade completeness score.

Table 3. The 5-year stratified summary statistics of well-documented Malaysian reports (N=56,877).

	Total, n (%)	2005-2009 (n=105), n (%)	2010-2014 (n=2843), n (%)	2015-2019 (n=53,929), n (%)	*P* value[a]
Sender type					<.001
Public specialist hospital	31,503 (55.4)	72 (68.6)	1542 (54.2)	29,889 (55.4)	
Public nonspecialist hospital	4954 (8.7)	3 (2.9)	217 (7.6)	4734 (8.8)	
Public clinic	15,609 (27.4)	4 (3.8)	866 (30.5)	14,739 (27.3)	
Other public services	3 (0)	0 (0)	1 (0)	2 (0)	
University hospital	496 (0.9)	9 (8.6)	25 (0.9)	462 (0.9)	
Private PRH[b]	939 (1.7)	11 (10.5)	58 (2)	870 (1.6)	
Private hospital or clinic	2114 (3.7)	6 (5.7)	106 (3.7)	2002 (3.7)	
Private community pharmacy	45 (0.1)	0 (0)	0 (0)	45 (0.1)	
Consumer	30 (0.1)	0 (0)	0 (0)	30 (0.1)	
Unknown	1184 (2.1)	0 (0)	28 (1)	1156 (2.1)	
Reporter qualification					<.001
Physician	11,446 (20.1)	53 (50.5)	488 (17.2)	10,905 (20.2)	
Pharmacist	40,295 (70.8)	16 (15.2)	1850 (65.1)	38,429 (71.3)	
Other HCP[c]	4084 (7.2)	36 (34.3)	500 (17.6)	3548 (6.6)	
Consumer	78 (0.1)	0 (0)	0 (0)	78 (0.1)	
Unknown	974 (1.7)	0 (0)	5 (0.2)	969 (1.8)	

[a]*P* value based on the Fisher exact test.

[b]PRH: product registration holder.

[c]HCP: health care professional.

Over the 15 years, most well-documented reports in Malaysia came from public health facilities, with public specialist hospitals contributing more than half (31,503/56,877, 55.38%). Public clinics emerged as key contributors in later stages, with well-documented reports increasing considerably from 3.8% (4/105) in the period from 2005 to 2009 to 30.46% (866/2843) in the following 5 years. Compared to public services, the private sector consistently demonstrated a marginal contribution to quality AE reporting in Malaysia. In the earlier years, physicians contributed approximately half (53/105, 50.5%) of the well-documented reports. In the subsequent periods, reports from pharmacists showed a rise in quantity and average completeness, yielding the highest overall rate of being well documented (1850/2843, 65.07% to 38,429/53,929, 71.26%) among all reporter types from 2010 to 2019.

Key Strategies and Interventions Implemented in Malaysia (2005-2019)

In Malaysia, various strategies and interventions were implemented over the 15 years with the intent of improving AE reporting, as summarized by 5-year period in Figure 4 [7,61-65]. While the impacts of most interventions are usually multifaceted at the national level and challenging to measure with limited quantitative information, there is a particular interest in understanding the influence of staffing levels at the PV center, the introduction of a new PV database, and enhancements to reporting tools on reporting quality at different time points.

Figure 4. Key strategies and interventions implemented to improve adverse event (AE) reporting in Malaysia between 2005 and 2019. CPD: continuing professional development; DIS: drug information service; HCP: health care professional; PhIS: pharmacy hospital information system; PRH: product registration holder; PRP: provisionally registered pharmacist; PV: pharmacovigilance; RiMUP: *risalah maklumat ubat untuk pengguna*.

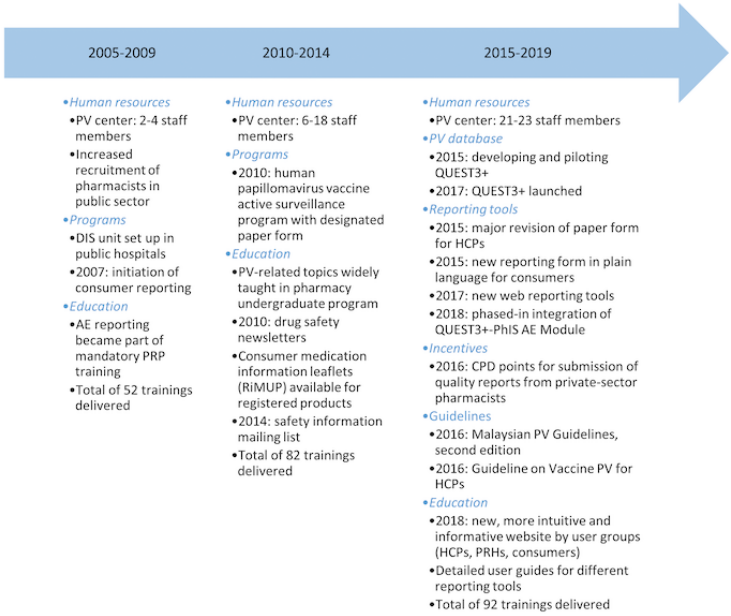


Figure 5 shows the annual trends in PV center staffing in relation to rates of reports being well documented. Figure 6 depicts how the transition in report submission format (from INTDIS to E2B) and reporting means correlated with report quantity and average completeness. In Figure 7, we focus on the rates of reports being well documented before and after the implementation of the new PV database (QUEST3+) and key enhancements to reporting tools since 2015. Information about reporting means was not available for INTDIS reports collected from the historical QUEST2 database. We further examined the influence and popularity of different reporting means among various reporters following the official launch of QUEST3+ and new web reporting tools in the first quarter of 2017 (Figure S4 in Multimedia Appendix 2).

Figure 5. Annual trends in pharmacovigilance (PV) center staffing levels and rate of reports being well documented (2005-2019).

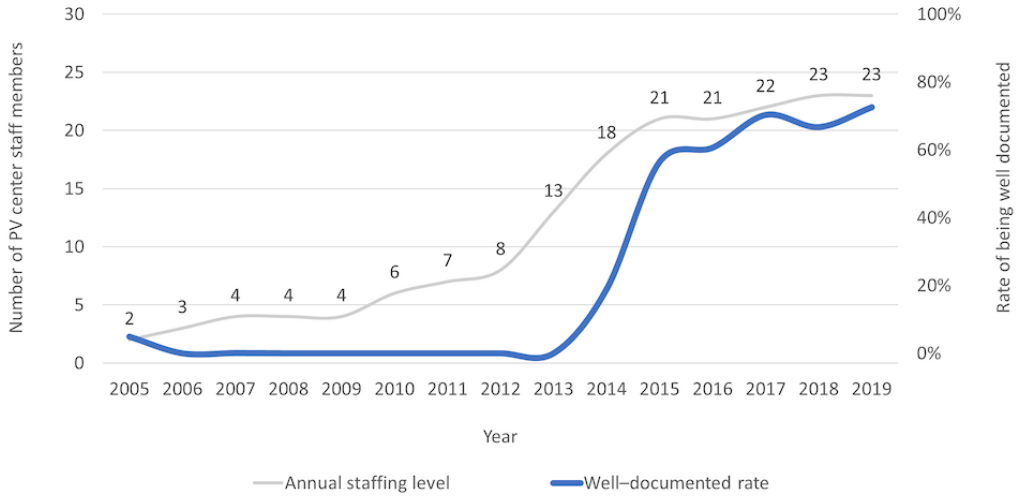


Figure 6. Mean completeness score and report count by submission format, pharmacovigilance database, and means of reporting yearly from 2013 to 2019. The size of the bubble corresponds to the report count. INTDIS: International Drug Information System; PhIS: pharmacy hospital information system.

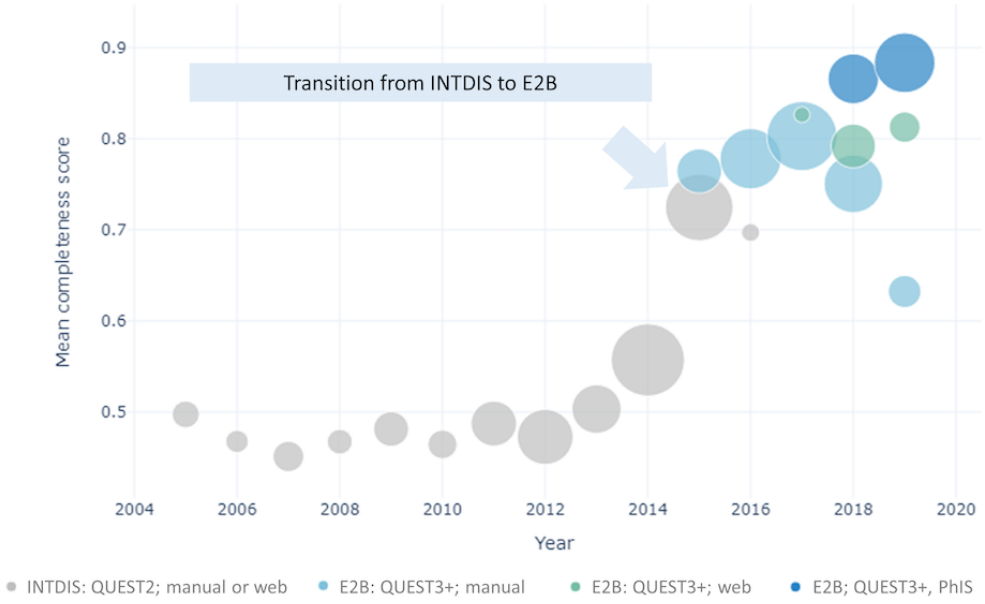
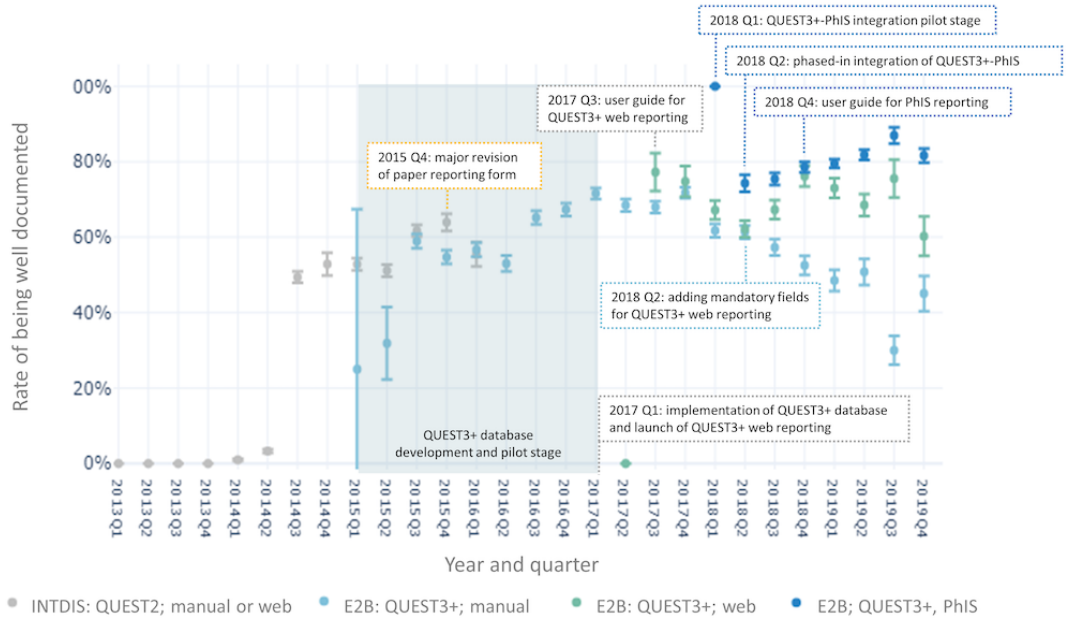


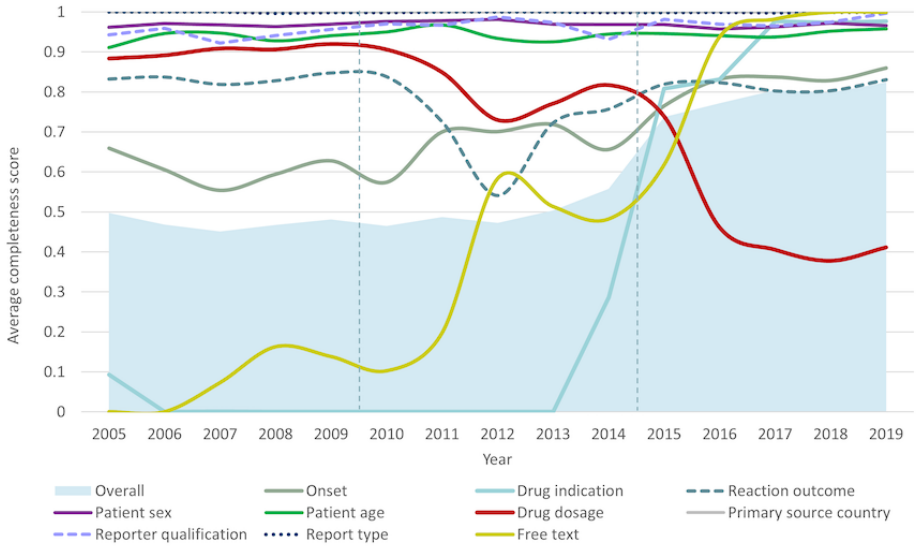
Figure 7. Rate of reports being well documented by submission format, pharmacovigilance database, and means of reporting quarterly from 2013 to 2019. 95% CI error bars (equivalent to 1.96×SE) were constructed. INTDIS: International Drug Information System; PhIS: pharmacy hospital information system; Q: quarter.



Average Completeness of Individual Dimensions for Malaysian AE Reports

Figure 8 illustrates the trends in the average completeness of the individual dimensions for Malaysian reports in VigiBase. The dimensions for report type, primary source country, reporter qualification, and patient age and sex were consistently the most completed. Completeness of free text and drug indication improved from zero in the earlier period of 2005 to 2009 to >0.9 in recent years. An uptrend in improvements was also observed for reaction onset. Completeness for reaction outcome dipped in 2011 to 2012 but subsequently rebounded to >0.8. Unexpectedly, we observed a noteworthy drop in drug dosage completeness since 2010. The average completeness of each individual dimension for vigiGrade was evaluated for E2B subsets (Figure S5 in Multimedia Appendix 2).

Figure 8. Trends in the distribution of average completeness scores for Malaysian reports in VigiBase over the study period. The shaded area indicates overall completeness; the line represents the individual dimension.



Discussion

Overview of Principal Findings

In our study, we used a comprehensive approach to examine the factors influencing the quality of Malaysian AE reports and decipher the temporal trends and interventions that shaped the Malaysian PV landscape from 2005 to 2019. In the first part of our study, by harnessing a data-driven approach that encompassed ML-based feature selection and analysis, we identified the key features that predict the status of Malaysian reports in VigiBase being well documented. Our hybrid feature selection helped in mitigating the risks of overfitting and unstable interpretability that are commonly associated with high variance [8,9,40] and resulted in a robust and valid RF model, as evidenced by the excellent classification performance (>90%) across all training, validation, and test sets for the E2B subset that reflects recent patterns of Malaysian reports (Table 1). While the model for the highly imbalanced INTDIS subset (containing only 10,691/63,943, 16.72% of well-documented reports) demonstrated satisfactory performance, with recall, precision, and F_1-scores of >70%, the model faced overfitting. This issue, evident from the decreased validation and test performance, has been acknowledged as a limitation in our study. The black-box RF model was made interpretable using SHAP values that, in agreement with human intuition, did not contradict the findings from the time-series analyses. To the best of our knowledge, this is the first study to use state-of-the-art interpretable ML methods to obtain insights concerning the key factors contributing to AE report quality in the PV field. Supplemented with insights drawn from our time-series and descriptive analyses, we summarized the identified features associated with well-documented Malaysian reports under 3 main themes: administrative; sender and reporter; and reaction, drug, and patient.

In the second part of our study, our extensive time-series and descriptive analyses illuminated the notable progress Malaysia has made in both the quantity and quality of AE reports over the years. We delved further into the chronological trends and characteristics of Malaysian AE reports, outlined the key strategies and interventions implemented at 5-year intervals, and tracked the influence of interventions of interest. These findings offer valuable insights into the multifaceted strategies and interventions that have driven enhancements in Malaysian AE reporting quality. Finally, by focusing on individual dimensions for vigiGrade completeness scoring, we not only gained additional insights into the specific characteristics and aspects of report completeness within the Malaysian context but also pinpointed systematic data quality issues that warrant further attention and improvement work.

Factors and Characteristics Associated With Well-Documented Malaysian Reports

Administrative

Our ML analysis distinguished PV center staffing level as the most important factor positively associated with well-documented INTDIS reports over 12 years between 2005 and 2016 (Figure S3 in Multimedia Appendix 2 and Figure 2A). Previous studies [66-68] have also highlighted manpower at national PV centers as a barrier to functional and sustainable PV systems. During the initial stage (2005-2009), only 2 to 4 staff members handled all PV operations in Malaysia (Figure 5). As the number of reports grew rapidly, the center struggled with a backlog of reports. The center swiftly grew from 8 to 21 staff members within 3 years (2012-2015) alongside the expansion of PV functions and task specializations, which coincided with a notable improvement in the completeness of reports received (Figures 3 and 5). Compared to the period from 2005 to 2009, more training workshops were delivered from 2010 to 2019 (Figure 4) to enhance staff and reporter competencies. As the staffing level and rate of reports being well documented became relatively stable afterward (Figure 5), the influence of PV staffing levels appeared less distinctive for E2B reports (Figure 2B). This could be due to some staff members focusing on other PV duties such as signal detection and assessment, risk management, and risk communication

rather than AE processing. It is worth noting that the transition to E2B may have also obfuscated the influence of increased staffing on report completeness. While less distinctive in our model, we cannot exclude that other centers with E2B reports and low staffing may still face overall low report completeness.

The capabilities and scalability of PV databases are essential for effective data collection and management [5]. Integrating electronic reporting into hospital information systems has been reported to reduce duplicate work and effectively improve AE reporting [69]. In Malaysia, the new QUEST3+ database was later integrated with the AE reporting module of a centralized PhIS implemented at Malaysian public hospitals and clinics, enabling the automatic input of reporter and sender information and improved reporter accessibility to patient, drug use, and regulatory information (eg, product registration number and batch number). Our ML analysis of E2B subsets (Figure 2B) revealed that PhIS-integrated reporting positively contributed to well-documented Malaysian reports, whereas web reporting had an unexpected negative association. The volume of PhIS reports surpassed other means in 2019 (Figure 7 and Figure S4 in Multimedia Appendix 2) and maintained the highest reporting quality. Web reporting was initially introduced in 2000, but due to unstable systems and slowly developing IT infrastructures at some public health facilities in the last decade, most reports before 2015 were submitted manually via postage, fax, or email. After mandatory fields were added, web reporting recorded a rate of >70% of reports being well documented, but it later declined to 60%. A similar but larger declining trend was also observed with manual reporting (Figure 5). These declines corresponded to the shift to PhIS reporting by public-sector pharmacists and were likely associated with other HCPs and consumers (Figure S4 in Multimedia Appendix 2). Our findings highlight that electronic reporting tools serve as ad hoc support for well-trained reporters but IT infrastructure maturity and widespread user acceptance are required for success. This is in line with a recent realist review [70] asserting that technological interventions alone, without capacity building, have little or no impact on health data quality.

Sender and Reporter

Our study revealed that the greatest proportion of well-documented Malaysian reports, amounting to 91.54% (52,066/56,877), originated from public health facilities overseen by the MOH (Table 3). Among the well-documented reports, 98.15% (55,825/56,877) were submitted by HCPs. This finding aligns with those of previous studies [71-73] conducted across different regions of Malaysia, which consistently highlight that most HCPs recognize AE reporting as part of their professional obligations. In Malaysia, consumer-generated reports constituted only 0.14% (78/56,877) of the well-documented reports, which stands in contrast to countries such as Denmark and Norway where three-fifths of well-documented reports came from consumers or non-HCPs [6].

Our observations revealed that Malaysian pharmacists generated 70.85% (40,295/56,877) of well-documented reports (Table 3), with most serving the public health sector [63]. Specifically, we identified public specialist hospitals and pharmacists as key features that positively contributed to the well-documented Malaysian E2B reports (Figure 2B). It is noteworthy that, during the earlier stage (2005-2009), pharmacists' reports exhibited the poorest quality and were recognized as a key factor predicting not well-documented INTDIS reports from 2005 to 2016 (Figure 2A). Nonetheless, over the years, following increased recruitment in public health services [74,75] and multifaceted initiatives aimed at strengthening pharmacists' roles and skills (Figure 4), pharmacists have become an integral part of AE monitoring in Malaysia. Almost 9 in 10 public hospital pharmacists in Malaysia have reported at least one AE in the past [76].

Malaysia presents a unique scenario in which pharmacists play a leading role in PV activities, distinguishing it from global trends, in which physicians contribute nearly two-thirds of well-documented reports [6]. The Malaysian context aligns with findings from a Spanish study in which pharmacists reported a great majority of the AEs due to the integration of PV into routine hospital pharmacy practices [77]. Similar observations were made in a pharmacist-led AE monitoring and management model in China, where pharmacists provided higher-quality reports among all HCPs [14]. Within Malaysian public hospitals, the pharmacist-led Drug Information Service (DIS) unit is responsible for facility-level PV activities, including responding to queries related to AEs; disseminating safety information; and compiling, verifying, and submitting AE reports to the NPRA [65,78]. In addition to direct detection and reporting by pharmacists, a collaborative mechanism exists within public health facilities where physicians and other HCPs are aware of the role of pharmacists in monitoring and reporting the AEs detected during clinical rounds or discussions. Moreover, we observed that over half (19,188/33,559, 57.18%) of the well-documented E2B reports made by pharmacists came from public specialist hospitals. These reports were believed to have benefited from the input of specialist physicians, suggesting a positive contribution of collaborative efforts among HCPs in enhancing the quality of AE reports.

In contrast, reports from PRHs and other HCPs (including regulatory affairs officers, clinical trial associates, nurses, and medical assistants) were flagged as the key features negatively associated with Malaysian report completeness (Figure 2B). These findings are consistent with the features observed in the United States [15], Brazil [16], Spain [17], South Korea [18], and Japan [19,20]. Of note, the NPRA classified the reports from PRHs as reported by other HCPs when the primary reporter was unknown. Among 7833 E2B reports from PRHs, only 778 (9.93%) reports were well documented (Table S1 in Multimedia Appendix 2), with an overall completeness score of 0.39 (Figure S5 in Multimedia Appendix 2). Information regarding drug dosage, reaction onset, reaction outcome, and patient age was most incomplete in Malaysian reports from PRHs. This could be attributed to the lack of a robust PV culture among PRHs. Existing literature [6,79] suggests that PRHs might prioritize submitting a report to fulfill pharmaceutical legislation [80] that mandates that PRHs report any suspected AEs within strict timelines even when minimal information is available. There could also be instances in which the primary reporter did not provide consent for follow-up. Conversely, it is conceivable that pharmacists and physicians serving at health facilities were

more motivated to make a clinically meaningful report even on a voluntary basis [17,79,81].

Reaction, Drug, and Patient

As AE reporting is highly dependent on individual motivation [5], we were interested in understanding whether the nature of drugs and reactions affects the quality of reports. While a report may involve more than one drug or reaction, most studies on AE report quality have not assessed all drugs and reactions reported in a case. Toki and Ono [21] examined only the primary suspected drug in a multivariable logistic regression model, whereas Araujo et al [22] and Masuka and Khoza [23] evaluated a specific drug group using simple univariable analysis. Other studies have evaluated only case-level information, such as case seriousness [14,18,24-26], fatal outcome [15], and causality [18]. As we converted drug-reaction pairs to case-level data, we evaluated the influence of drug- and reaction-related factors on overall report completeness for all reported suspected and interacting drugs.

Our ML analysis of the E2B subset (Figure 2B) revealed that a case where the reaction abated following a drug dechallenge (ie, positive dechallenge) was the primary key feature associated with a well-documented report. While information on drug dechallenge was unknown in most cases (Table S1 in Multimedia Appendix 2), it is possible that a positive dechallenge may have strengthened the reporter's confidence in the drug-reaction causal relationship and, thus, the motivation to construct a clinically meaningful report. While our findings suggest that positive dechallenge may have motivated more complete reports, there is no supporting study on this. It is important to note that reports that are well documented by vigiGrade standards might also tend to have a positive dechallenge. In other words, it may be that vigiGrade tends to flag those reports that have a positive dechallenge as well documented.

As expected, cases containing information on time to onset were more likely to be well documented as the onset dimension incurs the highest penalty of 50% for missing data and 30% if the uncertainty exceeds 1 month [6]. Our findings suggest that cases with a shorter (ie, <1 day) time to onset, dosing interval, and duration of drug use were most likely to be well documented. This observation might be attributable to better recall and description of events occurring within a shorter time frame following drug use or to greater reporter confidence in the drug-reaction relationship due to stronger temporal association and a lower likelihood of confounding factors. On the other hand, reports that involved reactions lasting 1 to 6 days tended to carry more information compared to those that involved reactions lasting <1 day or >6 days. A competing hypothesis is that reactions occurring within this time frame allow for sufficient time for more observation or data gathering while still being easily observed and described by patients and HCPs. Nevertheless, further research is needed to determine the specific factors contributing to the observed differences in report quality for cases with varying time to onset, dosing intervals, and durations of drug use.

Antibiotics for systemic use (ATC code J01) emerged as a key feature favorably contributing to well-documented E2B reports. First, it could be attributed to the baseline reporting patterns, where systemic antibiotics were the most commonly reported drug group in Malaysia, with over one-fifth of AE reports involving at least one systemic antibiotic (Table S1 in Multimedia Appendix 2). Second, this observation might suggest that Malaysian HCPs exercise heightened caution when using anti-infectives, leading to a higher likelihood of detecting AEs related to anti-infectives with higher report completeness. According to a study from a Malaysian infectious disease hospital [65], most inquiries (37.8%) received by the DIS unit concerned anti-infective drugs (ATC code J), which included other β-lactam antibacterials (ATC code J01D), direct-acting antivirals (ATC code J05A), and penicillins (ATC code J01C), with the largest proportion of the inquiries pertaining to their AEs and pediatric dosage adjustments. Although this observation could be expected in an infectious disease hospital, it also highlights the role of pharmacist-led DIS in AE monitoring and reinforces our previous discussion regarding pharmacists working in public health facilities tending to submit more complete reports. Trainings by the NPRA often prioritized pharmacists working in DIS units, who then conducted echo training for HCPs in their respective health facilities.

In addition, our analysis revealed a positive association between reports marked as serious and well-documented Malaysian reports (Figure 2B), consistent with previous studies from France [25,26], China [14,24], and South Korea [18] that highlighted higher completeness for serious reports. Previous research has also indicated that Malaysian HCPs prioritize reporting serious AEs [71]. The heightened gravity and potential consequences of these cases might prompt reporters to exercise greater diligence in ensuring reporting quality, including PRHs who are subjected to stricter reporting timelines for serious cases. Fatal outcomes were not flagged as the key feature contributing to Malaysian reports being well documented, likely due to their low prevalence (1048/68,795, 1.52%; Table S1 in Multimedia Appendix 2). Nonetheless, Malaysian fatal reports had a low overall completeness of 0.55 compared to 0.80 for nonfatal reports (Figure S5 in Multimedia Appendix 2). Another observational study evaluating reports submitted to the US Food and Drug Administration [15] also found that cases of patient deaths had the lowest completeness scores across reporting sources. This could be attributed to the absence of medical terminology describing the cause of death or indicate an investigation into a potential drug involvement.

Reactions under the HLGTs *product quality, supply, distribution, manufacturing,* and *quality system issues*, of which 97.78% (2242/2293) were related to product substitution issues and 91.41% (2096/2293) were sourced from public health facilities primarily by pharmacists, were captured as the key feature that negatively contributed to E2B reports being well documented. Subsequent investigations revealed that product substitution issues were most prevalent in 2018 (1355/2242, 60.44%) and primarily involved brand switching between 2 generic products: amlodipine (1012/2242, 45.14%) and perindopril (291/2242, 12.98%). Among these, only 24.8% (251/1012) and 32.6% (95/291) of the reports involving amlodipine and perindopril, respectively, were well documented. In the Malaysian public health care sector, drugs are procured

through 3 distinct mechanisms: a national concessionaire, national tenders, and direct purchases by health facilities for items not covered by the former 2 mechanisms [82]. However, in situations in which a product substitution issue is suspected and public health facilities need to directly procure alternative products for items already listed in the former 2 mechanisms, AEs or product complaints must be submitted as justification. It is believed that the reporters might submit reports containing only minimal information solely to comply with the drug procurement procedure.

Another key negative feature identified in the E2B subset was the presence of adolescent patients (aged 12-17 years). In comparison, reports involving adult patients (aged 18-44 years) and midlife adult patients (aged 45-64 years), which comprised the largest proportion of Malaysian reports, tended to be well documented. In South Korea, overall reports involving children and adolescents (aged 0-19 years) were negatively associated with being well documented in comparison to the older adult group (aged ≥65 years), whereas reports involving adults (aged 19-65 years) had a positive association [18].

Trends in Malaysian AE Reporting Quality Between 2005 and 2019

Building on the preceding discussion on the factors and characteristics associated with well-documented Malaysian AE reports, we expanded our scope to the chronological progression of AE reporting in Malaysia from 2005 to 2019. Our analyses underline that policy changes, continuity of education, and human resource development laid the foundation for a functional and sustainable SRS in Malaysia. Meanwhile, advancements in technological infrastructure, PV databases, and reporting tools contributed to the observed increase in both the quantity and quality of AE reports. These findings echo the expert-recommended 4-tier hierarchy of needs to achieve systemic capacity building for PV [83]—progressing from structures, systems, and roles to staff and infrastructure to skills to tools.

Malaysia, with its SRS governed within an established legal and regulatory framework, has historically struggled with challenges of underreporting and poorly reported AEs, as evidenced in Figure 3. In an effort to establish a functional and sustainable PV system, Malaysia placed early priorities on cultivating a reporting culture among HCPs and strengthening national PV capacities through collaborative efforts involving multiple stakeholders (Figure 4). Among them were policy changes to strengthen pharmacists' role in AE monitoring, increased recruitment of public-sector pharmacists and PV staff, active surveillance programs for targeted medicinal products, public awareness campaigns, and continuity of PV education to HCPs from undergraduate and preservice to at-service levels [7,61-63,65,74,75]. These initiatives were consistent with the existing literature, which emphasizes that multifaceted strategies and interventions work more synergistically to improve AE reporting than a single intervention [79,84,85]. Notably, our findings from the ML analysis suggest a positive association between higher staffing levels at the PV center and well-documented INTDIS reports, which could underscore the potential need for capacity building in the early phase of PV implementation.

As PV activities in Malaysia attained a higher level of maturation, the NPRA began to put greater emphasis on improving report quality. Comparative studies examining reporting forms from various countries have consistently highlighted that the Malaysian paper reporting form captures the most comprehensive information [86,87]. In response to the influx of reports observed in 2014 (Figure 3), the NPRA set their efforts on enhancing AE reporting tools and processing capabilities (Figure 4). Enhancements were made to the paper reporting form in 2015, including the addition of structured checkboxes and a reporting guide to ensure that more complete and harmonized clinical information could be obtained for subsequent causality assessment [27,88]. Concurrently, the NPRA began developing and piloting QUEST3+, an upgraded regulatory database system that marked a new submission format to the UMC—transitioning from INTDIS to E2B (Figure 6). The official launch of the QUEST3+ database took place in January 2017, replacing the historical QUEST2 database. Alongside these paradigm shifts, the NPRA also revamped and relaunched its web reporting tool for HCPs and introduced a new plain-language web reporting tool specifically for consumers (ConSERF). With the maturation of IT infrastructures, in 2018, the QUEST3+ database was integrated in phases with the centralized PhIS across Malaysian public health facilities. Reporting guides, drop-down lists, and validation alerts were also added to web and PhIS-integrated reporting tools to enhance the completeness and consistency of the collected data. Interestingly, as previously discussed, our comparative findings regarding PhIS-integrated tools used by well-trained pharmacists and new web reporting tools likely used by other HCPs and consumers highlight the complementary role of electronic reporting tools as ad hoc aids for well-trained reporters, whereas the effectiveness of these tools in improving AE reporting also relies on the maturity of IT infrastructures and their acceptance by users.

As a consequence of continuous efforts to strengthen PV capacities and technological advancements, Malaysia has seen considerable improvements not only in the quantity but also in the quality of reports. From 2015 to 2019, approximately two-thirds of Malaysian reports were well documented compared to approximately 1 in 5 reports from the rest of the world [28]. It is worth noting that, while overall completeness improved after the transition to the E2B submission format in 2015, our investigations revealed that low completeness in drug dosage (Figure 8) was systematically confounded by miscoding errors during report conversion to E2B-XML files before report transmission to VigiBase and, thus, was comparatively lowest in all subsets (Figure S5 in Multimedia Appendix 2). As a consequence of missing "number of unit in the interval" and miscoded "number of separate dosages," the drug dosage dimension for a Malaysian report in E2B format was penalized when the total daily dose for a case could not be calculated from the specified fields [29,89]. Similar to global reports [6], Malaysian E2B reports carried more administrative information, such as report type followed by reporter qualification and patient characteristics (ie, sex and age), but less drug- and

reaction-related information, such as drug dosage (despite the aforementioned confounding), reaction onset, and reaction outcome. In contrast to global reports [6], the inclusion of mandatory fields in electronic reporting tools led to a higher completeness of drug indication and free-text narratives in Malaysian reports. Reports from the literature and other sources, made by other HCPs, and submitted by PRHs had the lowest overall completeness scores (<0.5). Fatal reports and those from community pharmacies or other public services also tended to contain less information (<0.6).

Limitations

Our study is constrained by the limitations inherent to cross-sectional observational data and ML analysis, where causal reasoning and statistical inference cannot be determined [54]. The features identified in our study should be understood as predictors associated with well-documented reports but not as causal factors. Owing to the assumptions that multifaceted interventions often work synergistically and that control groups are frequently absent in nationwide implementation [79], the exact impact of individual interventions on reporting quality cannot be determined. As such, our study serves as an exploratory analysis, and the highlighted features offer a starting point for further in-depth review.

Our study faced challenges with the class imbalance inherent to the INTDIS data set, which heightened the risk of model overfitting. While undersampling improved the balanced performance of precision and recall, it could introduce new biases. Given that the INTDIS format is now obsolete in Malaysia, our focus is shifting toward the more recent E2B features.

Our models did not include causality information for several reasons. AE reports received by the NPRA and VigiBase come from a variety of sources, and the probability that the suspected AE is drug related is not the same in all cases [90]. Reporters and senders might use different methods for assessing causality, such as Naranjo probability scores and WHO-UMC causality categories, which were not available. In addition, it is important to note that causality may change as knowledge expands, and the UMC does not validate the causality assessments of the received reports.

Our data set is also constrained by the timeliness of report submission from QUEST to VigiBase and did not include all the reports received by the NPRA by December 31, 2019. As the systematic data quality issues uncovered in our study have already been communicated to the NPRA, follow-up work is underway to address these issues. Therefore, the findings of this study will not be representative of the future completeness scores of Malaysian reports in VigiBase. This also implies that the key features identified in our study were subject to multiple systematic biases, which are typically encountered when using real-world data [90,91].

Conclusions and Future Work

By using a data-driven approach and the vigiGrade method, we pinpointed the trends and milestones of the Malaysian AE reporting system and demonstrated how the country has striven to contribute large numbers of high-quality reports to global PV. Our work also highlights the vigiGrade method by the UMC as an effective tool for monitoring the quality of AE reports and aiding countries in evaluating to enhance reporting. Our multidimensional perspective on AE reporting trends and strategies in Malaysia, informed by data-driven insights, underlines the complexity and evolving nature of the SRS and the importance of continual improvement for global PV.

Using interpretable ML methods, we identified specific features that were positively associated with Malaysian AE reports being well documented. Notable factors include higher PV center staffing for INTDIS reports, reaction abated upon drug dechallenge, reaction onset or drug use duration of <1 week, dosing interval of <1 day, reports from public specialist hospitals, reports by pharmacists, and reaction duration between 1 and 6 days for recent E2B reports. Conversely, reports from PRHs and other HCPs indicated areas for potential improvement in the quality of Malaysian reports. These identified features could potentially serve as a basis for future research and strategies aimed at improving PV practices, thus improving drug safety surveillance and, ultimately, public health outcomes.

Furthermore, our time-series analysis showcased how Malaysia has built up and strengthened its PV capacity via multifaceted strategies and interventions to enhance both the quantity and quality of AE reports. Policy changes, continuity of education, and human resource development have all contributed to the foundation for a functional and sustainable SRS in Malaysia, whereas advancements in technological infrastructure, PV databases, and reporting tools concurred with the rise in both the quantity and quality of AE reports. These findings resonate with the expert-recommended 4-tier hierarchy of needs for systemic PV capacity building—from structures, systems, and roles to staff and infrastructure to skills to tools [83,92].

Building on our findings on Malaysia's progress in AE reporting and factors identified for report quality, we propose several areas for future work. To understand how and in what measure the findings from the time-series analysis contributed to the completeness of Malaysian reports, viewing the interventions set up by the NPRA as complex [93]—targeting multiple individuals or a wide range of behaviors and involving multiple interacting components—could be instrumental. Future evaluations may use this newly updated framework for complex intervention research [94].

Our findings revealed that the private health sector, including PRHs, private hospitals, private clinics, and community pharmacies, exhibited suboptimal contributions. This highlights persistent challenges pertaining to underreporting and unsatisfactory report quality in these sectors, necessitating further research into understanding behavioral or organizational barriers for developing targeted interventions [95,96]. Considering that preservice and in-service trainings often do not adequately prepare HCPs for data-related tasks [96], stronger stakeholder coordination and collaboration are imperative for continuous competency-based training and fostering an effective data use culture across health systems [95]. Regular feedback on reporting performance could be considered to facilitate self-monitoring among all senders.

Sustainable improvement in surveillance data quality and use requires a whole-systems approach encompassing governance, people, tools, and processes [95]. Given that data quality is highly reliant on their collection at health facilities, future work can prioritize people and environments essential for functional information systems as well as validation upon data entry to ensure completeness, accuracy, and consistency [88]. Our identification of systematic data quality issues highlighted a gap in data-driven continuous quality improvement [95], underscoring the need for internal quality assurance procedures for AE data management and transmission, including routine systematic checks and periodic in-depth reviews [88,97].

Looking ahead, as Malaysian reports currently use the E2B(R2) format, future efforts can navigate toward transitioning to the E2B(R3)-compliant database and reporting tools as the inclusion of null flavors in the E2B(R3) format helps address missing information by explaining data absence. Future work on data governance could explore leveraging automation, ML, and natural language processing to improve the overall efficiency and quality of AE data collection, processing, and management [98,99].

Acknowledgments

This paper is based on the master's thesis completed by SMC at the Graduate Institute of Biomedical Informatics, Taipei Medical University, under the supervision of SS-A and the joint supervision of DS and Jim Barrett from the Uppsala Monitoring Centre (UMC), as well as SCL from the National Pharmaceutical Regulatory Agency (NPRA). No generative artificial intelligence tools were used to create the original content for publication. The views expressed in this paper do not represent the opinions of the NPRA, the UMC, or the World Health Organization. The authors thank the Director-General of Health Malaysia for his permission to publish this study and are indebted to all stakeholders who contributed reports to QUEST and VigiBase. Special thanks to Usman Iqbal, Ekansh Gayakwad, Suo-Chen Chien, and Yu-Chin Chu from Taipei Medical University and Wai Lam Hoo from the University of Malaya for their valuable assistance and advice. The authors are also grateful for the unwavering support provided by Azuana Ramli, Norleen Mohamed Ali, Kobu Thiruvanackan, Nora Ashikin Mohd Ali, Mohd Ghazli Ismail, and Wee Kee Wo from the NPRA. The publication fund for this research was sponsored in part by the National Science and Technology Council (NSTC) Taiwan under grant NSTC 110-2320-B-038-029-MY3.

Data Availability

The data sets generated during and analyzed during this study are not publicly available due to the data protection policies of the Uppsala Monitoring Centre and the National Pharmaceutical Regulatory Agency. Data requests should be made directly to the institutions.

Authors' Contributions

SMC contributed to conceptualization, methodology, data curation, software, formal analysis, and original draft writing. DS and SCL contributed equally to conceptualization, methodology, formal analysis, and supervision. SS-A contributed to conceptualization, methodology, supervision, and funding acquisition. HCY contributed to methodology, formal analysis, and funding acquisition. All authors have reviewed, edited, and approved the final manuscript. SS-A and HCY are co-corresponding authors on this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of definitions or operational definitions.
[PDF File (Adobe PDF File), 243 KB - medinform_v12i1e49643_app1.pdf]

Multimedia Appendix 2
Supplementary figures and tables.
[PDF File (Adobe PDF File), 1866 KB - medinform_v12i1e49643_app2.pdf]

Multimedia Appendix 3
Literature review on quality of reports in the spontaneous reporting system.
[PDF File (Adobe PDF File), 101 KB - medinform_v12i1e49643_app3.pdf]

Multimedia Appendix 4
Feature selection.
[PDF File (Adobe PDF File), 206 KB - medinform_v12i1e49643_app4.pdf]

Multimedia Appendix 5
Tree-based machine learning models and interpretable machine learning.
[PDF File (Adobe PDF File), 79 KB - medinform_v12i1e49643_app5.pdf]

References

1. The importance of pharmacovigilance. World Health Organization. 2002. URL: https://apps.who.int/iris/handle/10665/42493 [accessed 2021-05-22]
2. Natsiavas P, Malousi A, Bousquet C, Jaulent MC, Koutkias V. Computational advances in drug safety: systematic and mapping review of knowledge engineering based approaches. Front Pharmacol 2019 May 17;10:415 [FREE Full text] [doi: 10.3389/fphar.2019.00415] [Medline: 31156424]
3. Bate A, Hornbuckle K, Juhaeri J, Motsko SP, Reynolds RF. Hypothesis-free signal detection in healthcare databases: finding its value for pharmacovigilance. Ther Adv Drug Saf 2019 Aug 5;10:2042098619864744 [FREE Full text] [doi: 10.1177/2042098619864744] [Medline: 31428307]
4. Uppsala Monitoring Centre. URL: https://www.who-umc.org/ [accessed 2021-05-22]
5. García CH, Pinheiro L, Maciá MA, Stroe R, Georgescu A, Dondera R, et al. Spontaneous adverse drug reactions: subgroup report. Heads of Medicines Agencies, European Medicines Agency. URL: https://tinyurl.com/4n76snn5 [accessed 2021-05-22]
6. Bergvall T, Norén GN, Lindquist M. vigiGrade: a tool to identify well-documented individual case reports and highlight systematic data quality issues. Drug Saf 2013 Dec 17;37(1):65-77. [doi: 10.1007/s40264-013-0131-x]
7. National Pharmaceutical Regulatory Agency, Ministry of Health Malaysia. URL: https://npra.gov.my/index.php/en/ [accessed 2021-02-22]
8. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health 2020 Apr 02;41:21-36 [FREE Full text] [doi: 10.1146/annurev-publhealth-040119-094437] [Medline: 31577910]
9. Stevens LM, Linstead E, Hall JL, Kao DP. Association between coffee intake and incident heart failure risk. Circ Heart Failure 2021 Feb;14(2):e006799. [doi: 10.1161/circheartfailure.119.006799]
10. Merriam-Webster, Inc. Merriam-Webster Dictionary. Merriam-Webster, Inc. URL: https://www.merriam-webster.com/ [accessed 2024-03-25]
11. Klein K, Scholl JH, De Bruin ML, van Puijenbroek EP, Leufkens HG, Stolk P. When more is less: an exploratory study of the precautionary reporting bias and its impact on safety signal detection. Clin Pharma Therapeutics 2017 Oct 25;103(2):296-303. [doi: 10.1002/cpt.879]
12. Edwards IR, Lindquist M, Wiholm BE, Napke E. Quality criteria for early signals of possible adverse drug reactions. The Lancet 1990 Jul;336(8708):156-158 [FREE Full text] [doi: 10.1016/0140-6736(90)91669-2]
13. Caster O, Juhlin K, Watson S, Norén GN. Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank. Drug Saf 2014 Jul 23;37(8):617-628. [doi: 10.1007/s40264-014-0204-5]
14. Chen Y, Niu R, Xiang Y, Wang N, Bai J, Feng B. The quality of spontaneous adverse drug reaction reports in China: a descriptive study. Biol Pharml Bull 2019;42(12):2083-2088. [doi: 10.1248/bpb.b19-00637]
15. Moore TJ, Furberg CD, Mattison DR, Cohen MR. Completeness of serious adverse drug event reports received by the US Food and Drug Administration in 2014. Pharmacoepidemiol Drug Safety 2016 Feb 10;25(6):713-718. [doi: 10.1002/pds.3979]
16. Ribeiro A, Lima S, Zampieri ME, Peinado M, Figueras A. Filling quality of the reports of adverse drug reactions received at the Pharmacovigilance Centre of São Paulo (Brazil): missing information hinders the analysis of suspected associations. Expert Opin Drug Saf 2017 Aug 23;16(12):1329-1334. [doi: 10.1080/14740338.2017.1369525]
17. Plessis L, Gómez A, García N, Cereza G, Figueras A. Lack of essential information in spontaneous reports of adverse drug reactions in Catalonia—a restraint to the potentiality for signal detection. Eur J Clin Pharmacol 2017 Mar 1;73(6):751-758. [doi: 10.1007/s00228-017-2223-5]
18. Oh IS, Baek YH, Kim HJ, Lee M, Shin JY. Differential completeness of spontaneous adverse event reports among hospitals/clinics, pharmacies, consumers, and pharmaceutical companies in South Korea. PLoS ONE 2019 Feb 14;14(2):e0212336. [doi: 10.1371/journal.pone.0212336]
19. Tsuchiya M, Obara T, Miyazaki M, Noda A, Sakai T, Funakoshi R, et al. High-quality reports and their characteristics in the Japanese Adverse Drug Event Report database (JADER). J Pharm Pharm Sci 2021 Apr 08;24:161-173. [doi: 10.18433/jpps31417]
20. Tsuchiya M, Obara T, Sakai T, Nomura K, Takamura C, Mano N. Quality evaluation of the Japanese Adverse Drug Event Report database (JADER). Pharmacoepidemiol Drug 2019 Dec 10;29(2):173-181. [doi: 10.1002/pds.4944]
21. Toki T, Ono S. Assessment of factors associated with completeness of spontaneous adverse event reporting in the United States: a comparison between consumer reports and healthcare professional reports. J Clin Pharm Ther 2019 Nov 25;45(3):462-469. [doi: 10.1111/jcpt.13086]
22. Araujo AG, Lucchetta RC, Tonin FS, Pontarolo R, Borba HH, Wiens A. Analysis of completeness for spontaneous reporting of disease-modifying therapies in multiple sclerosis. Expert Opin Drug Saf 2021 Mar 11;20(6):735-740. [doi: 10.1080/14740338.2021.1897566]
23. Masuka JT, Khoza S. An analysis of the trends, characteristics, scope, and performance of the Zimbabwean pharmacovigilance reporting scheme. Pharmacol Res Perspect 2020 Sep 15;8(5):e00657. [doi: 10.1002/prp2.657]

24. Niu R, Xiang Y, Wu T, Zhang Z, Chen Y, Feng B. The quality of spontaneous adverse drug reaction reports from the pharmacovigilance centre in western China. Expert Opin Drug Saf 2019 Jan;18(1):51-58. [doi: 10.1080/14740338.2019.1559812] [Medline: 30574811]
25. Humbert X, Jacquot J, Alexandre J, Sassier M, Robin N, Pageot C, et al. Completeness of pharmacovigilance reporting in general medicine in France. Sante Publique 2019;31(4):561-566. [doi: 10.3917/spub.194.0561]
26. Durrieu G, Jacquot J, Mège M, Bondon-Guitton E, Rousseau V, Montastruc F, et al. Completeness of spontaneous adverse drug reaction reports sent by general practitioners to a regional pharmacovigilance centre: a descriptive study. Drug Saf 2016 Sep 29;39(12):1189-1195. [doi: 10.1007/s40264-016-0463-4]
27. Bahk CY, Goshgarian M, Donahue K, Freifeld CC, Menone CM, Pierce CE, et al. Increasing patient engagement in pharmacovigilance through online community outreach and mobile reporting applications: an analysis of adverse event reporting for the Essure device in the US. Pharmaceut Med 2015;29(6):331-340 [FREE Full text] [doi: 10.1007/s40290-015-0106-6] [Medline: 26635479]
28. Wakao R, Taavola H, Sandberg L, Iwasa E, Soejima S, Chandler R, et al. Data-driven identification of adverse event reporting patterns for Japan in VigiBase, the WHO global database of individual case safety reports. Drug Saf 2019 Sep 26;42(12):1487-1498. [doi: 10.1007/s40264-019-00861-y]
29. Technical description of vigiGrade: completeness score. Uppsala Monitoring Centre. URL: https://tinyurl.com/2b55m5k8 [accessed 2024-03-11]
30. Kheloufi F, Default A, Rouby F, Laugier-Castellan D, Boyer M, Rodrigues B, et al. Informativeness of patient initial reports of adverse drug reactions. Can it be improved by a pharmacovigilance centre? Eur J Clin Pharmacol 2017 Aug;73(8):1009-1018. [doi: 10.1007/s00228-017-2254-y] [Medline: 28391408]
31. Muñoz MA, Delcher C, Dal Pan GJ, Kortepeter CM, Wu E, Wei YJ, et al. Impact of a new consumer form on the quantity and quality of adverse event reports submitted to the United States Food and Drug Administration. Pharmacotherapy 2019 Nov;39(11):1042-1052. [doi: 10.1002/phar.2325] [Medline: 31479525]
32. Fernandez-Fernandez C, Lázaro-Bengoa E, Fernández-Antón E, Quiroga-González L, Montero Corominas D. Quantity is not enough: completeness of suspected adverse drug reaction reports in Spain-differences between regional pharmacovigilance centres and pharmaceutical industry. Eur J Clin Pharmacol 2020 Aug;76(8):1175-1181. [doi: 10.1007/s00228-020-02894-0] [Medline: 32447435]
33. Tsuchiya M, Obara T, Miyazaki M, Noda A, Takamura C, Mano N. The quality assessment of the Japanese Adverse Drug Event Report database using vigiGrade. Int J Clin Pharm 2020 Apr;42(2):728-736. [doi: 10.1007/s11096-020-00969-7] [Medline: 32020439]
34. Rolfes L, van Hunsel F, van der Linden L, Taxis K, van Puijenbroek E. The quality of clinical information in adverse drug reaction reports by patients and healthcare professionals: a retrospective comparative analysis. Drug Saf 2017 Jul;40(7):607-614 [FREE Full text] [doi: 10.1007/s40264-017-0530-5] [Medline: 28405899]
35. Masuka JT, Khoza S. Adverse events following immunisation (AEFI) reports from the Zimbabwe expanded programme on immunisation (ZEPI): an analysis of spontaneous reports in Vigibase from 1997 to 2017. BMC Public Health 2019 Aug 27;19(1):1166 [FREE Full text] [doi: 10.1186/s12889-019-7482-x] [Medline: 31455314]
36. Uppsala reports 68. Uppsala Monitoring Centre. 2015 Jan. URL: https://who-umc.org/media/164371/ur68_final_2_gb.pdf [accessed 2024-03-12]
37. Oosterhuis I, Taavola H, Tregunno PM, Mas P, Gama S, Newbould V, et al. Characteristics, quality and contribution to signal detection of spontaneous reports of adverse drug reactions via the WEB-RADR mobile application: a descriptive cross-sectional study. Drug Saf 2018 Oct;41(10):969-978 [FREE Full text] [doi: 10.1007/s40264-018-0679-6] [Medline: 29761281]
38. Jokinen J, Bertin D, Donzanti B, Hormbrey J, Simmons V, Li H, et al. Industry assessment of the contribution of patient support programs, market research programs, and social media to patient safety. Ther Innov Regul Sci 2019 Nov;53(6):736-745. [doi: 10.1177/2168479019877384] [Medline: 31684774]
39. Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney Res Clin Pract 2017 Mar 31;36(1):3-11. [doi: 10.23876/j.krcp.2017.36.1.3]
40. Kuhn M, Johnson K. Feature Engineering and Selection: A Practical Approach for Predictive Models. Boca Raton, FL: CRC Press; 2019.
41. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007 Oct 01;23(19):2507-2517. [doi: 10.1093/bioinformatics/btm344] [Medline: 17720704]
42. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020 Oct;13(10) [FREE Full text] [doi: 10.1161/circoutcomes.120.006556]
43. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003 Mar 1;3:1157-1182 [FREE Full text] [doi: 10.1162/153244303322753616]
44. 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]

45. Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods 2018 Jun;15(6):399-400. [doi: 10.1038/s41592-018-0019-x] [Medline: 29855577]
46. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods 2016 Aug 30;13(9):703-704. [doi: 10.1038/nmeth.3968]
47. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997 Dec;97(1-2):273-324. [doi: 10.1016/S0004-3702(97)00043-X]
48. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform 2018 Aug;116:10-17 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.05.006] [Medline: 29887230]
49. Lundberg SM, 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; December 4-9, 2017; Long Beach, CA.
50. van den Bosch T, Warps AL, de Nerée tot Babberich MP, Stamm C, Geerts BF, Vermeulen L, et al. Predictors of 30-day mortality among Dutch patients undergoing colorectal cancer surgery, 2011-2016. JAMA Netw Open 2021 Apr 26;4(4):e217737. [doi: 10.1001/jamanetworkopen.2021.7737]
51. Gong K, Lee HK, Yu K, Xie X, Li J. A prediction and interpretation framework of acute kidney injury in critical care. J Biomed Inform 2021 Jan;113:103653 [FREE Full text] [doi: 10.1016/j.jbi.2020.103653]
52. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
53. Ariza-Garzon MJ, Arroyo J, Caparrini A, Segovia-Vargas MJ. Explainability of a machine learning granting scoring model in peer-to-peer lending. IEEE Access 2020;8:64873-64890. [doi: 10.1109/access.2020.2984412]
54. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci U S A 2019 Oct 16;116(44):22071-22080. [doi: 10.1073/pnas.1900654116]
55. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Victoria, CA: Leanpub; 2020.
56. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health 2020 Apr;2(4):e179-e191. [doi: 10.1016/s2589-7500(20)30018-2]
57. Shapley L. A value for n-person games. In: Kuhn H, Tucker A, editors. Contributions to the Theory of Games II. Princeton, NJ: Princeton University Press; 1953:307-317.
58. Breiman L. Random forests. Mach Learn 2001;45:5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
59. Seabold S, Perktold J. Statsmodels econometric and modeling with python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SciPy 2010; June 28-July 3, 2010; Austin, TX. [doi: 10.25080/majora-92bf1922-011]
60. 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 [FREE Full text]
61. National centre for adverse drug reactions monitoring annual report. National Pharmaceutical Regulatory Agency (NPRA), Ministry of Health Malaysia. 2021 Nov 29. URL: https://tinyurl.com/44ayxzwk [accessed 2022-02-22]
62. Elkalmi RM, Hassali MA, Al-lela OQ, Jamshed SQ. The teaching of subjects related to pharmacovigilance in Malaysian pharmacy undergraduate programs. J Pharmacovigilance 2013;1(2):1-5. [doi: 10.4172/2329-6887.1000106]
63. Annual Report and Statistics of the Pharmaceutical Services Programme. Pharmaceutical Services Programme, Ministry of Health Malaysia. 2019. URL: https://tinyurl.com/y4h7ctn4 [accessed 2021-05-22]
64. Rosli R, Dali AF, Aziz NA, Ming LC, Manan MM. Reported adverse drug reactions in infants: a nationwide analysis in Malaysia. Front Pharmacol 2017 Feb 10;8:30 [FREE Full text] [doi: 10.3389/fphar.2017.00030] [Medline: 28239351]
65. Ali A, Mohd Yusoff S, Mohd Joffry S, Wahab MS. Drug information service awareness program and its impact on characteristics of inquiries at DIS unit in Malaysian public hospital. Arch Pharma Pract 2013;4(1):9. [doi: 10.4103/2045-080x.111576]
66. Suwankesawong W, Dhippayom T, Tan-Koi WC, Kongkaew C. Pharmacovigilance activities in ASEAN countries. Pharmacoepidemiol Drug Saf 2016 Sep 12;25(9):1061-1069. [doi: 10.1002/pds.4023] [Medline: 27174034]
67. Ampadu HH, Hoekman J, Arhinful D, Amoama-Dapaah M, Leufkens HG, Dodoo AN. Organizational capacities of national pharmacovigilance centres in Africa: assessment of resource elements associated with successful and unsuccessful pharmacovigilance experiences. Global Health 2018 Nov 16;14(1):109 [FREE Full text] [doi: 10.1186/s12992-018-0431-0] [Medline: 30445979]
68. Olsson S, Pal SN, Stergachis A, Couper M. Pharmacovigilance activities in 55 low- and middle-income countries: a questionnaire-based analysis. Drug Saf 2010 Aug 01;33(8):689-703. [doi: 10.2165/11536390-000000000-00000] [Medline: 20635827]
69. Ortega A, Aguinagalde A, Lacasa C, Aquerreta I, Fernández-Benítez M, Fernández LM. Efficacy of an adverse drug reaction electronic reporting system integrated into a hospital information system. Ann Pharmacother 2008 Aug 26;42(10):1491-1496. [doi: 10.1345/aph.1l130]

70. A realist review of what works to improve data use for immunization: evidence from low- and middle-income countries. Pan American Health Organization, World Health Organization. 2019. URL: https://tinyurl.com/3p5dveue [accessed 2024-03-04]
71. Balan S. Knowledge, attitude and practice of Malaysian healthcare professionals toward adverse drug reaction reporting: a systematic review. Int J Pharm Pract 2021 Aug 11;29(4):308-320. [doi: 10.1093/ijpp/riab030] [Medline: 34289016]
72. Ali RS, Ismail WI. Adverse drug reactions reporting: knowledge, attitude and practice among healthcare providers at a tertiary hospital in northern region of Malaysia. Asian J Med Health Sci 2021 Oct;4(Supplement 1):214-227 [FREE Full text]
73. Kirthikaa GK. Evaluation of knowledge, attitude, and practice towards adverse drug reaction reporting and reason for underreporting among the private and public medical practitioners of Kuala Lumpur and Selangor. International Medical University Central Digital Repository. 2022. URL: https://rep.imu.edu.my/xmlui/handle/1234.56789/2945?show=full [accessed 2024-03-04]
74. Rosli R, Ming LC, Abd Aziz N, Manan MM. A retrospective analysis of spontaneous adverse drug reactions reports relating to paediatric patients. PLoS ONE 2016 Jun 1;11(6):e0155385. [doi: 10.1371/journal.pone.0155385]
75. Hadi MA, Ming LC. Impact of pharmacist recruitment on ADR reporting: Malaysian experience. South Med Rev 2011 Dec;4(2):102-103 [FREE Full text] [doi: 10.5655/smr.v4i2.1009] [Medline: 23093890]
76. Hadi MA, Helwani R, Long CM. Facilitators and barriers towards adverse drug reaction reporting: perspective of Malaysian hospital pharmacists. J Pharm Health Serv Res 2013 May 24;4(3):155-158. [doi: 10.1111/jphs.12022]
77. Pérez-Ricart A, Gea-Rodríguez E, Roca-Montañana A, Gil-Máñez E, Pérez-Feliu A. Integrating pharmacovigilance into the routine of pharmacy department: experience of nine years. Farm Hosp 2019 Jul 01;43(4):128-133. [doi: 10.7399/fh.11169] [Medline: 31276445]
78. Malayasian adverse drug reactions newsletter. National Pharmaceutical Control Bureau, Ministry of Health Malaysia. 2013 Apr. URL: https://tinyurl.com/3ywrmb5b [accessed 2021-05-22]
79. Li R, Zaidi ST, Chen T, Castelino R. Effectiveness of interventions to improve adverse drug reaction reporting by healthcare professionals over the last decade: a systematic review. Pharmacoepidemiol Drug Saf 2020 Jan;29(1):1-8 [FREE Full text] [doi: 10.1002/pds.4906] [Medline: 31724270]
80. Malaysian guidelines on Good Pharmacovigilance Pratices (GVP) for product registration holders. National Pharmaceutical Regulatory Agency (NPRA), Ministry of Health Malaysia. 2021 Sep 30. URL: https://tinyurl.com/mpass52p [accessed 2022-02-22]
81. Ribeiro-Vaz I, Silva AM, Costa Santos C, Cruz-Correia R. How to promote adverse drug reaction reports using information systems - a systematic review and meta-analysis. BMC Med Inform Decis Mak 2016 Mar 01;16:27 [FREE Full text] [doi: 10.1186/s12911-016-0265-8] [Medline: 26926375]
82. Hamzah NM, Perera PN, Rannan-Eliya RP. How well does Malaysia achieve value for money in public sector purchasing of medicines? Evidence from medicines procurement prices from 2010 to 2014. BMC Health Serv Res 2020 Jun 05;20:509. [doi: 10.1186/s12913-020-05362-8]
83. Potter C. Systemic capacity building: a hierarchy of needs. Health Policy Plan 2004 Sep 01;19(5):336-345. [doi: 10.1093/heapol/czh038]
84. Khalili M, Mesgarpour B, Sharifi H, Daneshvar Dehnavi S, Haghdoost AA. Interventions to improve adverse drug reaction reporting: a scoping review. Pharmacoepidemiol Drug 2020 May 19;29(9):965-992 [FREE Full text] [doi: 10.1002/pds.4966]
85. Gonzalez-Gonzalez C, Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Strategies to improve adverse drug reaction reporting: a critical and systematic review. Drug Saf 2013 May;36(5):317-328. [doi: 10.1007/s40264-013-0058-2] [Medline: 23640659]
86. Singh A, Bhatt P. Comparative evaluation of adverse drug reaction reporting forms for introduction of a spontaneous generic ADR form. J Pharmacol Pharmacotherapeutics 2022 Apr 11;3(3):228-232. [doi: 10.4103/0976-500x.99417]
87. Bandekar MS, Anwikar SR, Kshirsagar NA. Quality check of spontaneous adverse drug reaction reporting forms of different countries. Pharmacoepidemiol Drug 2010 Sep 15;19(11):1181-1185. [doi: 10.1002/pds.2004]
88. Increasing Adverse Event Reporting (IAER) subproject: survey report. International Coalition of Medicines Regulatory Authorities. URL: https://tinyurl.com/2ksfrm6z [accessed 2021-05-22]
89. Electronic transmission of individual case safety reports message specification. The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. 2001. URL: https://admin.ich.org/sites/default/files/inline-files/ICH_ICSR_Specification_V2-3.pdf [accessed 2024-03-12]
90. Guideline for using VigiBase data in studies. Uppsala Monitoring Centre (UMC). 2021 Mar 15. URL: https://who-umc.org/media/05kldqpj/guidelineusingvigibaseinstudies.pdf [accessed 2021-05-22]
91. 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]
92. Indicator-based pharmacovigilance assessment tool: manual for conducting assessments in developing countries. United States Agency for International Development. 2009 Dec. URL: https://pdf.usaid.gov/pdf_docs/pnads167.pdf [accessed 2021-05-22]
93. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337:a1655. [doi: 10.1136/bmj.a1655]

94. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061. [doi: 10.1136/bmj.n2061]
95. Scobie HM, Edelstein M, Nicol E, Morice A, Rahimi N, MacDonald NE, et al. Improving the quality and use of immunization and surveillance data: summary report of the Working Group of the Strategic Advisory Group of Experts on Immunization. Vaccine 2020 Oct;38(46):7183-7197. [doi: 10.1016/j.vaccine.2020.09.017]
96. Nicol E, Turawa E, Bonsu G. Pre- and in-service training of health care workers on immunization data management in LMICs: a scoping review. Hum Resour Health 2019 Dec 02;17:92. [doi: 10.1186/s12960-019-0437-6]
97. Radecka A, Loughlin L, Foy M, de Ferraz Guimaraes MV, Sarinic VM, Di Giusti MD, et al. Enhancing pharmacovigilance capabilities in the EU regulatory network: the SCOPE joint action. Drug Saf 2018 Aug 21;41(12):1285-1302 [FREE Full text] [doi: 10.1007/s40264-018-0708-5] [Medline: 30128638]
98. Ghosh R, Kempf D, Pufko A, Barrios Martinez LF, Davis CM, Sethi S. Automation opportunities in pharmacovigilance: an industry survey. Pharm Med 2020 Feb 08;34(1):7-18. [doi: 10.1007/s40290-019-00320-0]
99. Schmider J, Kumar K, LaForest C, Swankoski B, Naim K, Caubel PM. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. Clin Pharma Ther 2018 Dec 11;105(4):954-961. [doi: 10.1002/cpt.1255]

Abbreviations

AE: adverse event
ATC: Anatomical Therapeutic Chemical
DCA: Drug Control Authority
DIS: Drug Information Service
HCP: health care professional
HLGTs: High-Level Group Terms
ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
ICSR: individual case safety report
INTDIS: International Drug Information System
MADRAC: Malaysian Adverse Drug Reactions Advisory Committee
MedDRA: Medical Dictionary for Regulatory Activities
ML: machine learning
MOH: Ministry of Health
NPRA: National Pharmaceutical Regulatory Agency
PhIS: pharmacy hospital information system
PRH: product registration holder
PV: pharmacovigilance
RF: random forest
SHAP: Shapley additive explanations
SRS: spontaneous reporting system
UMC: Uppsala Monitoring Centre
WHO PIDM: World Health Organization Programme for International Drug Monitoring
WHO: World Health Organization

Edited by C Lovis; submitted 05.06.23; peer-reviewed by S Matsuda, C Zhao, I Degen; comments to author 20.08.23; revised version received 10.10.23; accepted 24.02.24; published 03.04.24.

Please cite as:
Choo SM, Sartori D, Lee SC, Yang HC, Syed-Abdul S
Data-Driven Identification of Factors That Influence the Quality of Adverse Event Reports: 15-Year Interpretable Machine Learning and Time-Series Analyses of VigiBase and QUEST
JMIR Med Inform 2024;12:e49643
URL: https://medinform.jmir.org/2024/1/e49643
doi:10.2196/49643
PMID:38568722

Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study

Ippei Akiya[1], MSc; Takuma Ishihara[2], PhD; Keiichi Yamamoto[3], PhD

1

2

3

Corresponding Author:
Ippei Akiya, MSc

Abstract

Background: Synthetic patient data (SPD) generation for survival analysis in oncology trials holds significant potential for accelerating clinical development. Various machine learning methods, including classification and regression trees (CART), random forest (RF), Bayesian network (BN), and conditional tabular generative adversarial network (CTGAN), have been used for this purpose, but their performance in reflecting actual patient survival data remains under investigation.

Objective: The aim of this study was to determine the most suitable SPD generation method for oncology trials, specifically focusing on both progression-free survival (PFS) and overall survival (OS), which are the primary evaluation end points in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods, including CART, RF, BN, and the CTGAN, and the performance of each method was evaluated.

Methods: Using multiple clinical trial data sets, 1000 data sets were generated by using each method for each clinical trial data set and evaluated as follows: (1) median survival time (MST) of PFS and OS; (2) hazard ratio distance (HRD), which indicates the similarity between the actual survival function and a synthetic survival function; and (3) visual analysis of Kaplan-Meier (KM) plots. Each method's ability to mimic the statistical properties of real patient data was evaluated from these multiple angles.

Results: In most simulation cases, CART demonstrated the high percentages of MSTs for synthetic data falling within the 95% CI range of the MST of the actual data. These percentages ranged from 88.8% to 98.0% for PFS and from 60.8% to 96.1% for OS. In the evaluation of HRD, CART revealed that HRD values were concentrated at approximately 0.9. Conversely, for the other methods, no consistent trend was observed for either PFS or OS. CART demonstrated better similarity than RF, in that CART caused overfitting and RF (a kind of ensemble learning approach) prevented it. In SPD generation, the statistical properties close to the actual data should be the focus, not a well-generalized prediction model. Both the BN and CTGAN methods cannot accurately reflect the statistical properties of the actual data because small data sets are not suitable.

Conclusions: As a method for generating SPD for survival data from small data sets, such as clinical trial data, CART demonstrated to be the most effective method compared to RF, BN, and CTGAN. Additionally, it is possible to improve CART-based generation methods by incorporating feature engineering and other methods in future work.

(JMIR Med Inform 2024;12:e55118) doi:10.2196/55118

KEYWORDS

oncology clinical trial; survival analysis; synthetic patient data; machine learning; SPD; simulation

Introduction

When submitting an application for the approval of a new pharmaceutical product to health authorities, it is imperative to demonstrate its efficacy and safety through multiple clinical trials. However, 86% of these trials encounter difficulties meeting the targeted number of subjects within the designated recruitment period, often leading to extensions of the trial duration or completion of the trial without reaching the target number of subjects [1-3]. The challenge of patient recruitment not only delays the submission of regulatory applications but also hinders the timely provision of effective treatment to patients, which consequently contributes to increased development costs and the escalation of drug prices and potentially exacerbates the strain on health care financing.

In recent years, the use of real-world data (RWD) has emerged as a potential solution for addressing these issues. The Food and Drug Administration has also released draft guidelines [4], garnering attention on the application of RWD as an external control arm in clinical trials [5,6]. Furthermore, it has been reported that it is possible to optimize eligibility using RWD and machine learning, thereby increasing the number of eligible subjects that can be included [7].

In addition to these approaches, we hypothesize that it is possible to generate synthetic patient data (SPD) from control arm data in past clinical trials and use it to establish a control arm for a new clinical trial. The use of SPD, an emerging research approach in the health care research field [8-17], involves the generation of fictitious individual patient-level data from real data, which possess statistical properties similar to those of actual data. This approach is anticipated to facilitate health care research while addressing data privacy concerns [14,18-21].

Regarding its application in clinical trials, concerns have been raised about the feasibility of generating SPDs with statistical properties similar to those of actual data due to the relatively smaller volume of clinical trial data compared to RWD, such as electronic health records or registry data. However, previous studies [22-25] have reported the successful generation of SPDs with statistical properties generally comparable to the actual data, although there are certain limitations. Additionally, with the expansion of clinical trial data-sharing platforms such as ClinicalStudyDataRequest.com, Project Data Sphere, and Vivli, acquiring subject-level clinical trial data has become more accessible. Consequently, advancements in research on the utility of SPD and the expansion of clinical trial data-sharing platforms are expected to have potential applications in clinical trials.

Our focus lies in the application of this technology in oncology clinical trials that evaluate popular efficacy end points such as overall survival (OS) and progression-free survival (PFS)–related survival functions and median survival time (MST) [26]. In previous studies on SPD, there has been a notable emphasis on reporting patient background data and single–time point data [22-25]. However, research focusing specifically on the relationship between SPD and survival data remains relatively insufficient [27].

As the first step in examining our hypothesis that the use of SPD can be beneficial in accelerating health care research, the aim of this study was to determine the most suitable SPD generation method for oncology trials, specifically focusing on both OS and PFS, which are set as the primary evaluation end points in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods: classification and regression trees (CART) [28], random forest (RF) [29], Bayesian network (BN) [30], and the conditional tabular generative adversarial network (CTGAN) approach [31], and the performance of each method was evaluated.

Methods

Overview

To generate the SPD, subject-level clinical trial data were obtained from Project Data Sphere for the following 4 clinical trials (Table 1): (1) each had a different cancer type, (2) included control arm data, (3) contained both OS and PFS data, and (4) had a ready data format for analysis.

Table . List of selected oncology clinical trials in this study.

ClinicalTrials.gov ID	Titles	Phase	Cancer type	Intervention for the control arm	Subjects in the control arm, n
NCT00119613	A Randomized, Double-Blind, Placebo-Controlled Study of Subjects With Previously Untreated Extensive-Stage Small-Cell Lung Cancer (SCLC) Treated With Platinum Plus Etoposide Chemotherapy With or Without Darbepoetin Alfa.	III	Small cell lung cancer	Placebo	232
NCT00339183	A Randomized, Multicenter Phase 3 Study to Compare the Efficacy of Panitumumab in Combination With Chemotherapy to the Efficacy of Chemotherapy Alone in Patients With Previously Treated Metastatic Colorectal Cancer.	III	Metastatic colorectal cancer	FOLFIRI[a] Alone	476
NCT00339183	A Phase 3 Randomized Trial of Chemotherapy With or Without Panitumumab in Patients With Metastatic and/or Recurrent Squamous Cell Carcinoma of the Head and Neck (SCCHN).	III	Recurrent or metastatic (or both) head and neck cancer	Cisplatin and 5-fluorouracil	260
NCT00703326	A Multicenter, Multinational, Randomized, Double-Blind, Phase III Study of IMC-1121B Plus Docetaxel versus Placebo Plus Docetaxel in Previously Untreated Patients With HER2-Negative, Unresectable, Locally-Recurrent or Metastatic Breast Cancer.	III	Breast cancer	Placebo and docetaxel	382

[a]FOLFIRI: panitumumab plus fluorouracil, leucovorin, and irinotecan.

Preparation of the Training Data Set

The patient data for the control arm contained within each trial data set were extracted and used as the actual data for the training data set. The selection of variables in the training data set aimed to include as many variables related to the subjects' background as possible, excluding variables concerning tests and evaluations conducted during the trials. Furthermore, variables that had the same value were excluded, even if they were related to the subjects' background (Multimedia Appendices 1-4).

Generation of Synthetic Data

The SPDs in this study were generated using the following 4 methods:

1. CART: the synthpop package (version 1.8) in R (The R Foundation) was used, specifying the cart method for the syn function's method argument.
2. RF: the synthpop package (version 1.8) in R was used, specifying the Ranger method for the syn function's method argument.
3. BN: the bnlearn package (version 4.9) in R was used to conduct structural learning through the score-based algorithm hill-climbing, followed by parameter estimation using the bn.fit function. The default maximum likelihood estimator was used for parameter estimation.
4. CTGAN: the CTGANSynthesizer module included in the Python package sdv (version 1.3) was used.

In all these generation methods, to ensure the absence of conflicting data regarding the relationship between PFS and OS, constraints were set to ensure that the values of PFS and OS were greater than zero and that PFS was less than or equal to OS. Specific individual patient data in the generated SPD, which did not meet these constraints, were excluded, and new individual patient data were regenerated. The SPDs were generated in a manner that equaled the number of subject-level data to the record count in the actual data.

To ensure the reproducibility of SPD generation, 1000 random numbers were generated as seed values using the Mersenne Twister algorithm. The same seed value set was used for all generation methods.

Statistical Analysis

Histogram

Histograms were created to visually inspect the distributions of the MST of the synthetic data (MSTS) for PFS and OS for the 1000 SPD data sets generated by each method. The histograms also included the MST of the actual data (MSTA) as a vertical line and the range of its 95% CI as a rectangular background. For PFS and OS, a higher percentage of MSTS covered by the 95% CI of the MSTA was determined to indicate a greater level of reliability for the generation method.

Evaluation of Similarity

A hazard ratio (HR) of 1 signifies that the 2 survival functions are entirely identical. Thus, the closer the HR is to 1, the more similar the 2 survival functions are. Accordingly, based on the following calculation formula, the HR distance (HRD) for PFS and OS from the SPD and the actual data were computed and evaluated:

$$HRD=1-abs(HR-1)$$

Kaplan-Meier Plot

In the evaluation of similarity, the SPD that showed the highest HRD value was considered the best case, and the SPD with the lowest HRD value was considered the worst case. Three groups of Kaplan-Meier (KM) plots were created, including the actual data, the best case, and the worst case for each SPD generation method. The best case and worst case for each SPD generation method in both PFS and OS were compared to actual survival by using the log rank test. Multiple comparisons were not performed, nor were *P* values adjusted because controlling for the type I error rate does not affect the conclusions of this study.

Since the purpose of this study was to evaluate the method of generating SPD that closely resemble actual survival data, it might be unnecessary to calculate a *P* value that indicates a significant difference from actual survival, but the *P* value was calculated in this study from the viewpoint that if a significant difference is also observed in the best-case, that method should not be adopted.

All analyses and data generation were performed using R (version 4.3.1; The R Foundation) and Python (version 3.10; Python Software Foundation).

Ethical Considerations

Ethical review was not needed for this simulation study for methodology comparison. All actual clinical trial data sets obtained from Project Data Sphere were used in accordance with relevant guidelines and regulations when the clinical trials were conducted.

Results

Figure 1 shows a histogram of the MSTS for PFS in the NCT00703326 trial. Using CART, RF, and BN, most of the generated MSTS values were within the 95% CI of the MSTA. In contrast, when CTGAN was used, SPD generation resulted in a widened variance in the distribution of MSTS. For the MSTS of PFS in the other 3 trials, RF exhibited a shift in the distribution of the MSTS, shortening the survival period, while BN displayed a shift in the distribution and prolonged the survival period. Similar trends to Figure 1 were observed for CART and CTGAN (Multimedia Appendices 5-7).

Figure 2 displays a histogram of the MSTS for OS in the NCT00460265 trial. The divergence from the PFS findings is that the MSTS of RF was more frequently included within the 95% CI of the MSTA, with similar results observed in other trials (Multimedia Appendices 8-10). In other aspects, similar findings were obtained as with the PFS.

Table 2 presents the number and proportion of the generated MSTS values included within the 95% CI of the MSTA for each trial and each method. In the case of CART for PFS, a high percentage ranging from 88.8% to 98.1% was exhibited for all trials. However, the OS ranged from 60.8% to 96.1%, with some trials displaying a lower percentage than the PFS.

Figure 1. Histogram of the median survival time of the synthetic data for progression-free survival in the NCT00703326 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; MST: median survival time; RF: random forest.

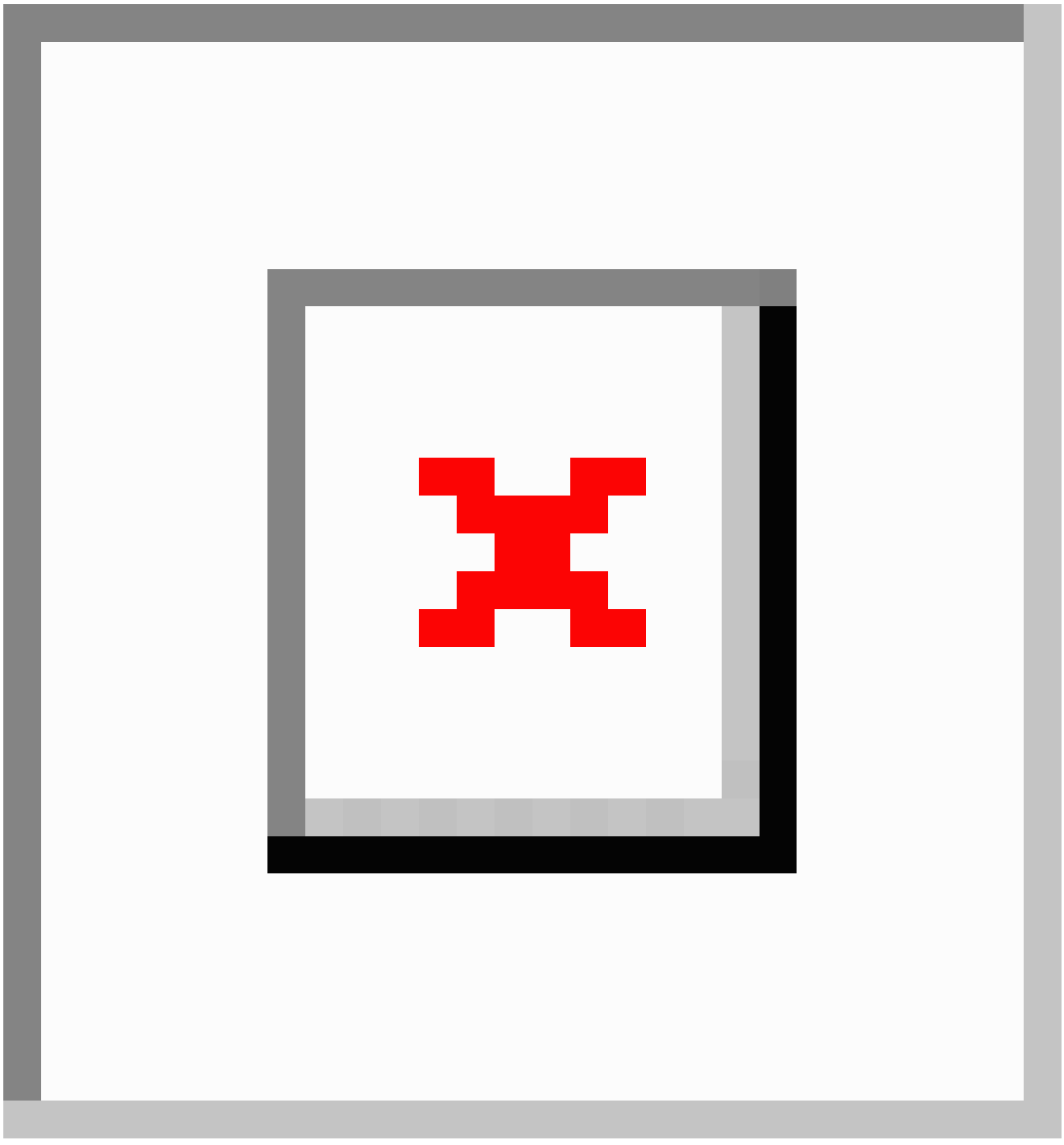


Figure 2. Histogram of the median survival time of the synthetic data of overall survival in the NCT00460265 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; MST: median survival time; RF: random forest.

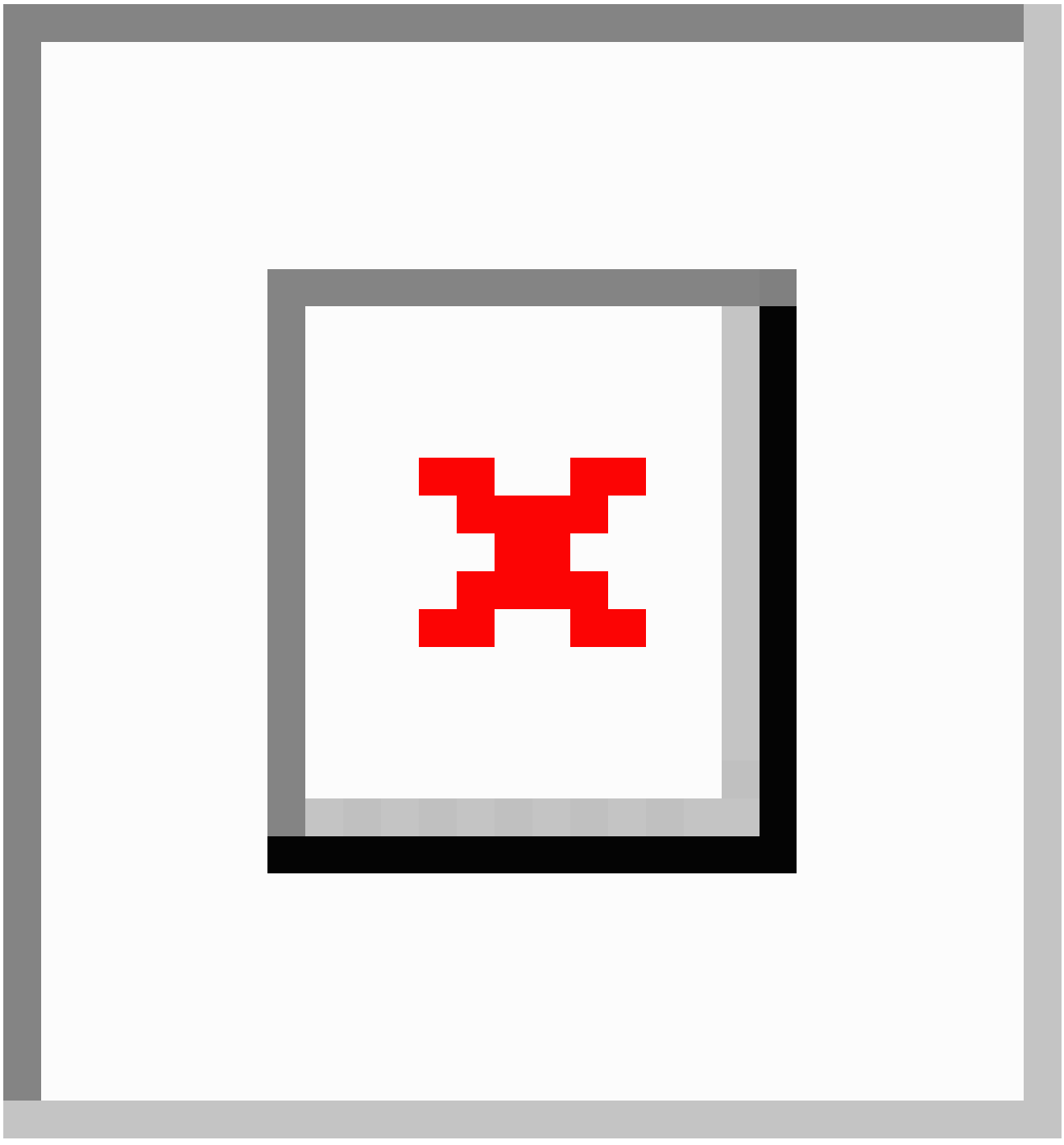


Table . The number and proportion of median survival times of the synthetic data (MSTSs) falling within the 95% CI of the median survival time of the actual data (MSTA).

		ClinicalTrials.gov ID			
		NCT00119613	NCT00339183	NCT00460265	NCT00703326
Progression-free survival					
	MSTA (95% CI)	169 (163-183)	155 (121-168)	133 (121-167)	424 (380-504)
MSTSs, n (%)					
	CART[a] (n=1000)	981 (98.1)	888 (88.8)	955 (95.5)	918 (91.8)
	RF[b] (n=1000)	693 (69.3)	248 (24.8)	426 (42.6)	919 (91.9)
	BN[c] (n=1000)	10 (1.0)	0 (0.0)	37 (3.7)	976 (97.6)
	CTGAN[d] (n=1000)	65 (6.5)	378 (37.8)	322 (32.2)	254 (25.5)
Overall survival					
	MSTA (95% CI)	276 (259-303)	361 (319-393)	286 (255-357)	1452 (1417-1507)
MSTSs, n (%)					
	CART (n=1000)	831 (83.1)	608 (60.8)	719 (71.9)	961 (96.1)
	RF (n=1000)	757 (75.7)	697 (69.7)	980 (98.0)	599 (59.9)
	BN (n=1000)	0 (0.0)	0 (0.0)	0 (0.0)	622 (62.2)
	CTGAN (n=1000)	72 (7.2)	155 (15.5)	197 (19.7)	81 (8.5)

[a]CART: classification and regression tree.

[b]RF: random forest.

[c]BN: Bayesian network.

[d]CTGAN: conditional tabular generative adversarial network.

For RF, a high proportion of 91.9% was observed for PFS in the NCT00703326 trial and 98.0% for OS in the NCT00460265 trial, whereas in other cases, the proportion for RF was not as high as that for CART.

In the case of BN, proportions of 97.6% and 62.2% were observed for PFS and OS, respectively, in the NCT00703326 trial, but in the other 3 trials, BN showed an extremely low percentage ranging from proportion ranging from 0.0% to 3.7%.

CTGAN showed a low proportion ranging from 6.5% to 37.8% for both PFS and OS in all trials.

Figure 3 shows the KM plot for PFS in the NCT00703326 trial. The best-case curves of CART and RF were similar to the actual data curve. In contrast, for BN and CTGAN, even the best-case curves deviated from the actual data curve. In other trials, some SPD did not show a similar trend. However, at least for the best-case scenarios of CART and RF, the generated synthetic survival curves closely resembled the actual survival curve (Multimedia Appendices 11-13).

Figure 4 displays the KM plot for OS in the NCT00460265 trial. Similar to the KM plots for PFS, the best-case curves of CART and RF resembled the actual data curve, whereas those of BN and CTGAN deviated from the actual data curve. These trends were also observed in other trials (Multimedia Appendices 14-16).

Figures 5 and 6 present box plots of the HRD. When using CART, the HRD values for both PFS and OS in all trials were concentrated at approximately 0.9. Conversely, for the other methods, no consistent trend was observed for either PFS or OS.

Figure 3. Kaplan-Meier plots for progression-free survival in the NCT00703326 trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

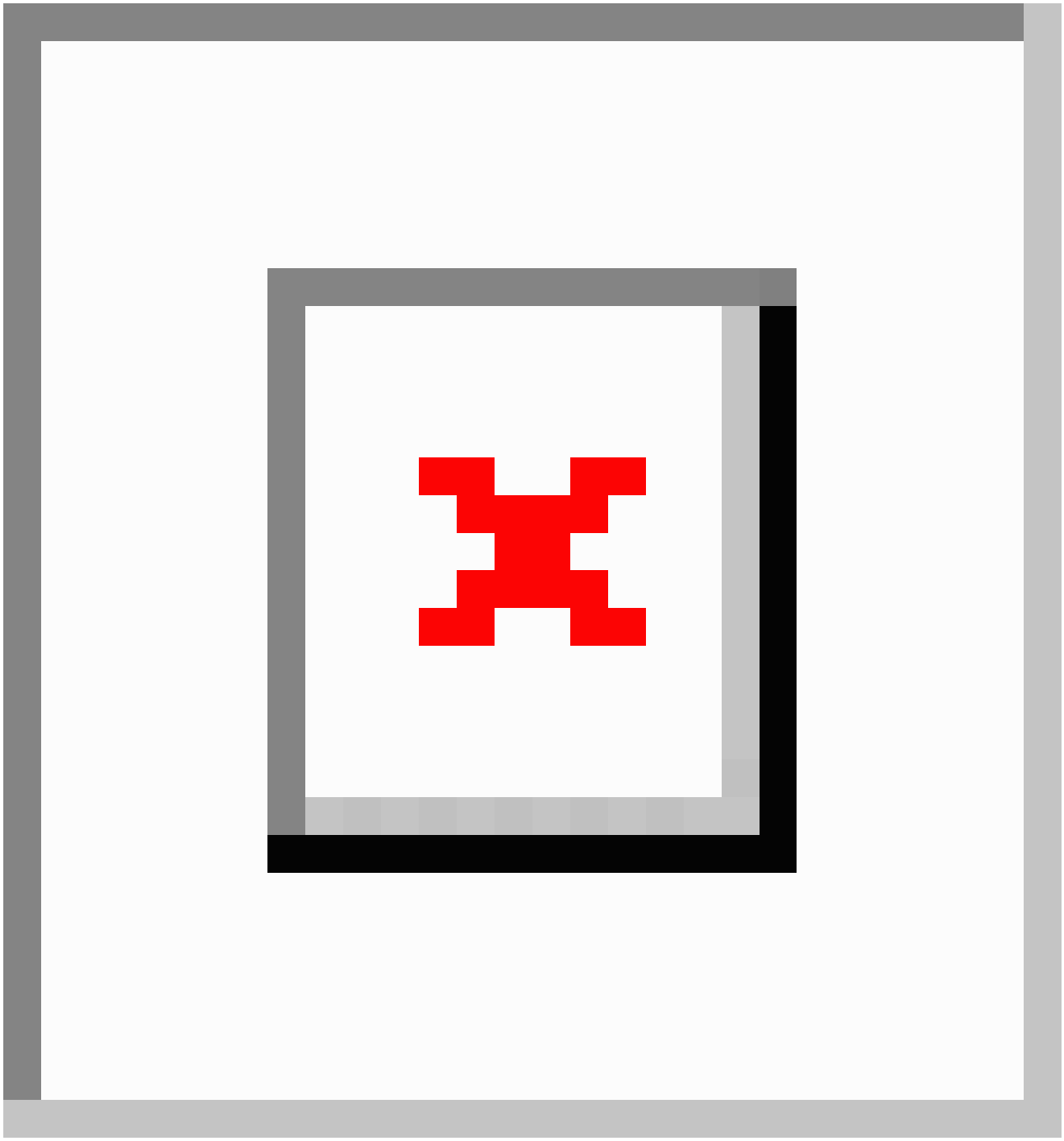


Figure 4. Kaplan-Meier plots for overall survival in the NCT00460265 trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

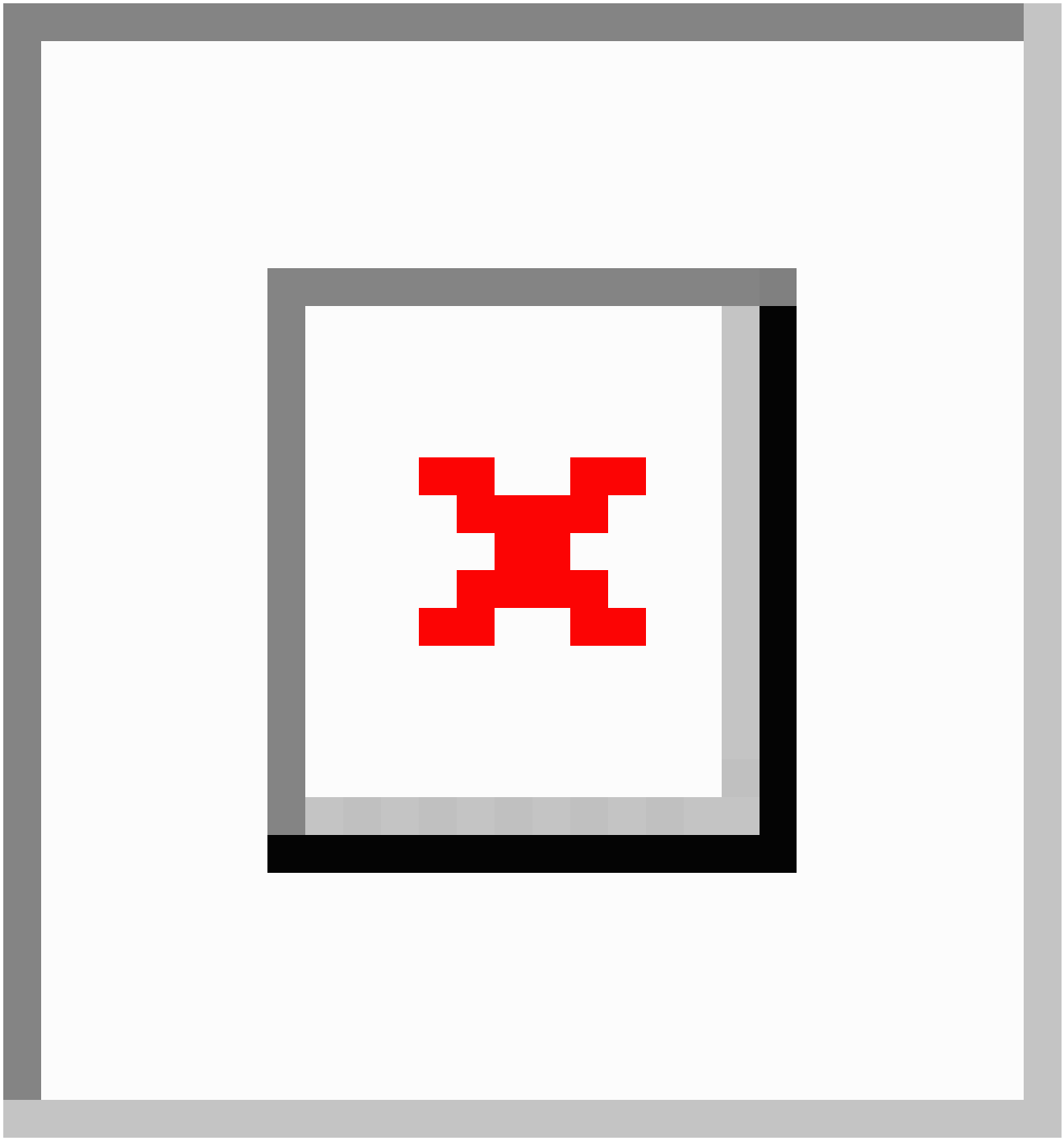


Figure 5. Box plot of progression-free survival hazard ratio distance (HRD) for each method and clinical trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

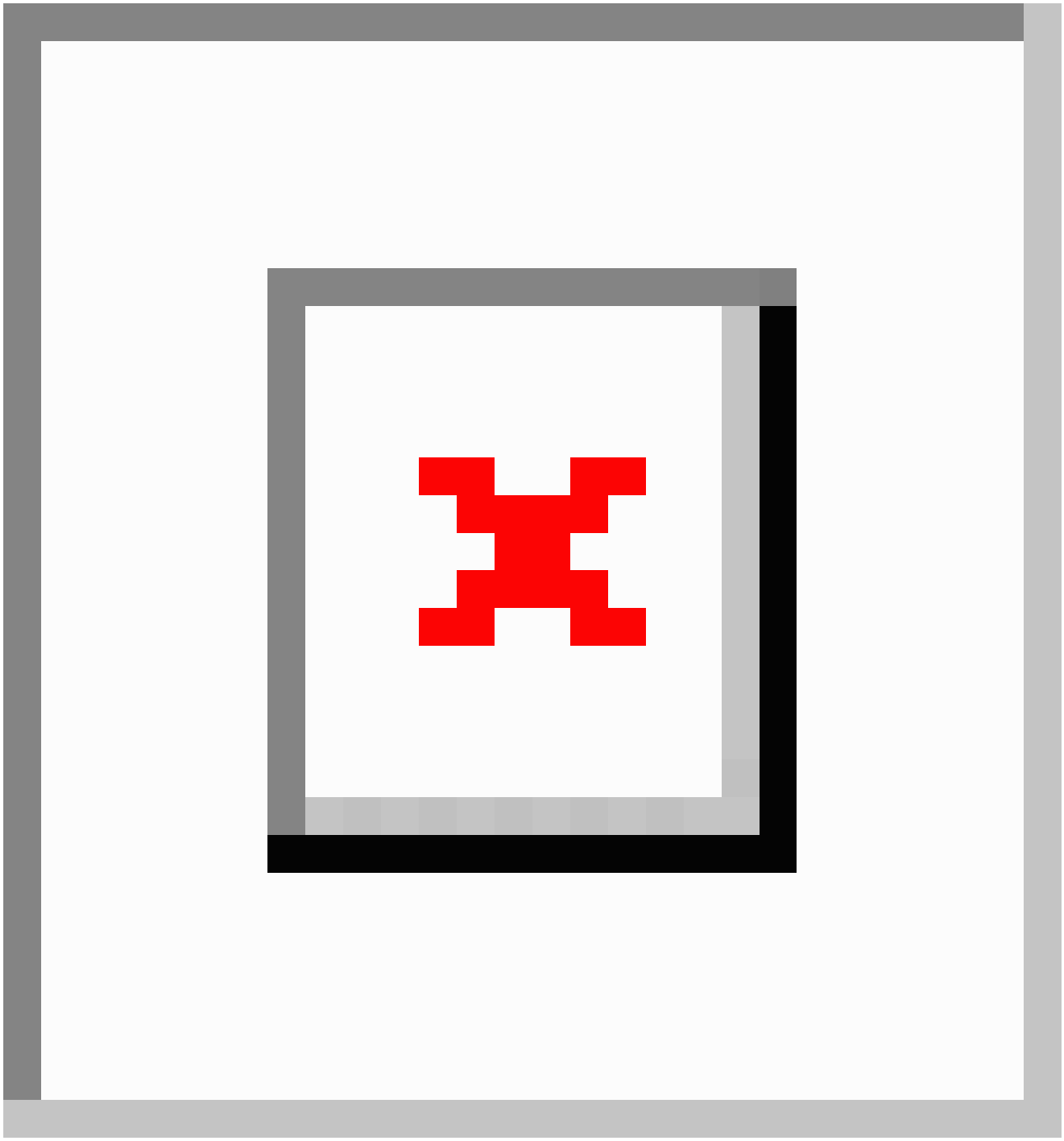
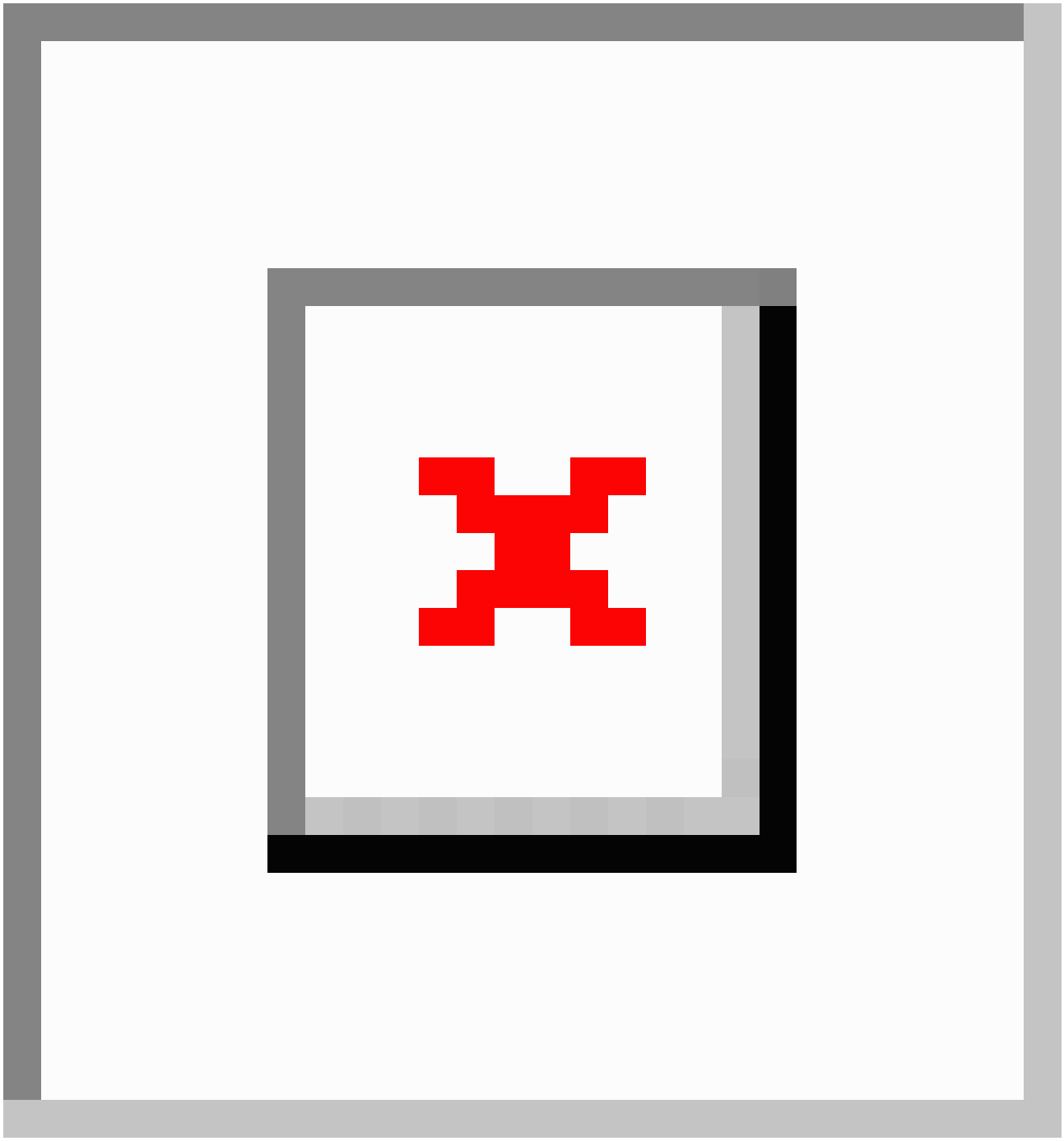


Figure 6. Box plot of overall survival hazard ratio distance (HRD) for each method and clinical trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.



Discussion

Regarding the survival SPD, CART often yielded better results than the other methods in evaluations using MST, HRD, and visual analysis via KM plots. Given the crucial importance of the hazard ratio and MST as end points in oncology trials [26], demonstrating the utility of both of these evaluation metrics is essential. Therefore, using CART for generating survival SPD was suggested as a beneficial approach.

While both CART and RF generally yielded preferable results across all trials, they share the common characteristic of using tree models. RF, with its use of the bootstrap method for resampling and constructing tree models for ensemble learning, is known to prevent overfitting. In general, in terms of constructing machine learning models with high generalization performance, RF performs better than CART. However, CART is prone to overfitting as the layers of the tree become deeper [32]. Although RF is considered a superior method for constructing high–generalization-performance machine learning models, the results from Table 2 and the KM plots in this study suggest that CART is a better approach than RF. This discrepancy might be due to differing views on what is a higher performance between the machine learning prediction model and SPD. In the machine learning prediction model, it is important to prevent overfitting and reduce bias; however, SPD is expected to match its statistical properties with actual data.

Thus, in the case of SPD, the overfitting suppression mechanism possessed by RF might have resulted in inferiority to that of CART from the perspective of improving similarity.

In the case of using BN, the percentage of MSTSs falling within the 95% CI of MSTAs was 0% for the PFS of the NCT00339183 trial, and for OS, this phenomenon also occurred in the NCT00119613, NCT00339183, and NCT0046265 trials. This implies that the SPD failed to accurately reflect the statistical properties of the actual data. Conversely, a high value of 97.6% was observed for the PFS in the NCT00703326 trial. The reason for this discrepancy could not be determined on the basis of the results of this study. Tucker et al [24] reported that they could generate data highly similar to actual data when using BN for the generation of SPD, which differs from our findings. One notable difference is that while Tucker et al [24] used a large-scale actual data set of 27.5 million patients for their study, this study used only a few hundred patients for training data. This difference likely had a significant impact on the accuracy of the SPD generation model, resulting in conflicting results. However, the SPD generated by BN were not distributed in the direction of shortening PFS or OS. Thus, this would not be harmful when the SPD generated by BN is used as a more conservative control arm in clinical trials.

Using CTGAN, the percentage of the MSTSs falling within the 95% CI of the actual data was low, indicating low performance associated with the generation of SPD that reflect the statistical properties of the actual data. However, Krenmayr et al [23] reported favorable performance results when using the same generative adversarial network (GAN)–based methods and RWD. The differences between their study and our study were as follows: their study did not include SPD on survival time or generate multiple SPD data sets from the same actual data, and there was a large amount of individual patient data in their study. In particular, focusing on the amount of individual patient data, the number of patients in each trial included in this study was relatively small, with the NCT00119613 trial having 232 patients, the NCT00339183 trial having 476 patients, the NCT0046265 trial having 260 patients, and the NCT00703326 trial having 382 patients, while the trial conducted by Krenmayr et al [23] had 500 or more patients. GAN-based methods using deep neural networks are known to perform poorly with small amounts of data [25,33]. In this study, although the NCT00339183 trial had the largest number of individual patient data, the best case of CTGAN for NCT00339183 produced a KM plot similar to the actual data, suggesting that a larger data set yields better results. Thus, there is no contradiction. Another characteristic of using CTGAN in this study was the larger variance in the estimated MSTSs, as indicated in Figures 1 and 2. Goncalves et al [34] showed that using MC-MedGAN, a GAN-based method, to generate an SPD from small data resulted in a large SD of the data utility metrics, leading to results with larger variance, similar to those of this study. Therefore, it is extremely challenging to generate useful SPD by applying GAN-based methods to small data sets, such as clinical trial data.

When generating SPDs for survival data and using them as a certain arm in a clinical trial, it is important to verify that the statistical properties closely match those of the actual data with the MST and the hazard ratio with the actual data being close to 1. Based on our results, we conclude that CART, which can concentrate the MSTSs within the range of 95% CI of MSTAs and approximately 0.9 for HRD, is an efficient method for generating SPD that meets the abovementioned conditions. However, even when using CART, slight variations were observed in the MSTSs, and some cases fell outside the 95% CI of the MSTAs, as revealed by our results. Therefore, for practical use, it is necessary to verify that the MSTSs are included in the 95% CI of the MSTAs and that both are close in value. It is also necessary to verify whether the HRD of the actual data and the SPD are close to 1 and then decide whether to adopt the generated SPD. Hence, the generation process must be repeated until an acceptable SPD is obtained. There may also be a need to use statistical methods to match characteristics between the SPD and the actual treatment arm in clinical trials.

In this study, even the most useful CART method produced SPDs that did not meet the requirements of MST and HRD. We expect that this issue will be addressed by incorporating feature engineering, such as dimension reduction, imputing missing values, derived variable creation, and other processing. Additionally, in clinical research, as subgroup analyses are frequently conducted, it is necessary to improve the generation method to reflect the statistical properties of the actual data even when the data are divided into subgroups under certain conditions. Moreover, from the perspective of data privacy, it is essential to incorporate approaches to prevent data reidentification into the generation method [35].

In conclusion, as a method for generating SPD for survival data from small data sets, such as clinical trial data, CART is the most effective method for generating SPD that meet the 2 conditions of having an MSTSs close to the MSTAs and an HRD close to 1. However, as SPD might be generated, which do not meet these 2 conditions, it is necessary to incorporate mechanisms to improve a CART-based generation method in future studies. Overcoming these challenges would make it possible to reduce the recruitment period and costs of clinical trial participants to ≥50% in comparative trials of new drug development against existing therapeutic drugs. This approach could accelerate clinical development, similar to the use of RWD.

Acknowledgments

We would like to express our gratitude to Project Data Sphere, the platform that provided the necessary data for this study, and to the clinical trial data providers Amgen and Eli Lilly.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variables used to generate synthetic patient data from the NCT00119613 trial.
[DOCX File, 48 KB - medinform_v12i1e55118_app1.docx]

Multimedia Appendix 2
Variables used to generate synthetic patient data from the NCT00339183 trial.
[DOCX File, 48 KB - medinform_v12i1e55118_app2.docx]

Multimedia Appendix 3
Variables used to generate synthetic patient data from the NCT00460265 trial.
[DOCX File, 47 KB - medinform_v12i1e55118_app3.docx]

Multimedia Appendix 4
Variables used for generating synthetic patient data from the NCT00703326 trial.
[DOCX File, 48 KB - medinform_v12i1e55118_app4.docx]

Multimedia Appendix 5
Histogram of the median survival time of the synthetic data for progression-free survival in the NCT00119613 trial. The dashed vertical line represents the median survival time for the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 202 KB - medinform_v12i1e55118_app5.docx]

Multimedia Appendix 6
Histogram of the median survival times for the synthetic data for progression-free survival in the NCT00339183 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 192 KB - medinform_v12i1e55118_app6.docx]

Multimedia Appendix 7
Histogram of the median survival times of the synthetic data for progression-free survival in the NCT00460265 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 187 KB - medinform_v12i1e55118_app7.docx]

Multimedia Appendix 8
Histogram of the median survival times of the synthetic data for overall survival in the NCT00119613 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 186 KB - medinform_v12i1e55118_app8.docx]

Multimedia Appendix 9
Histogram of the median survival times of the synthetic data for overall survival in the NCT00339183 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 195 KB - medinform_v12i1e55118_app9.docx]

Multimedia Appendix 10
Histogram of the median survival times of the synthetic data for overall survival in the NCT00703326 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File, 188 KB - medinform_v12i1e55118_app10.docx]

Multimedia Appendix 11
Kaplan-Meier plots for progression-free survival in the NCT00119613 trial.
[DOCX File, 215 KB - medinform_v12i1e55118_app11.docx]

Multimedia Appendix 12
Kaplan-Meier plots for progression-free survival in the NCT00339183 trial.

[DOCX File, 229 KB - medinform_v12i1e55118_app12.docx]

Multimedia Appendix 13
Kaplan-Meier plots for progression-free survival in the NCT00460265 trial.
[DOCX File, 218 KB - medinform_v12i1e55118_app13.docx]

Multimedia Appendix 14
Kaplan-Meier plots for overall survival in the NCT00119613 trial.
[DOCX File, 229 KB - medinform_v12i1e55118_app14.docx]

Multimedia Appendix 15
Kaplan-Meier plots for overall survival in the NCT00339183 trial.
[DOCX File, 252 KB - medinform_v12i1e55118_app15.docx]

Multimedia Appendix 16
Kaplan-Meier plots for overall survival in the NCT00703326 trial.
[DOCX File, 265 KB - medinform_v12i1e55118_app16.docx]

References

1. Huang GD, Bull J, Johnston McKee K, et al. Clinical trials recruitment planning: a proposed framework from the clinical trials transformation initiative. Contemp Clin Trials 2018 Mar;66:74-79. [doi: 10.1016/j.cct.2018.01.003] [Medline: 29330082]
2. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018 Sep;11:156-164. [doi: 10.1016/j.conctc.2018.08.001] [Medline: 30112460]
3. Treweek S, Lockhart P, Pitkethly M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open 2013;3(2):e002360. [doi: 10.1136/bmjopen-2012-002360] [Medline: 23396504]
4. Considerations for the design and conduct of externally controlled trials for drug and biological products. Guidance for industry. US Food and Drug Administration. 2023. URL: https://www.fda.gov/media/164960/download [accessed 2024-06-04]
5. Yap TA, Jacobs I, Baumfeld Andre E, Lee LJ, Beaupre D, Azoulay L. Application of real-world data to external control groups in oncology clinical trial drug development. Front Oncol 2021;11:695936. [doi: 10.3389/fonc.2021.695936] [Medline: 35070951]
6. Dagenais S, Russo L, Madsen A, Webster J, Becnel L. Use of real‐world evidence to drive drug development strategy and inform clinical trial design. Clin Pharmacol Ther 2022 Jan;111(1):77-89. [doi: 10.1002/cpt.2480] [Medline: 34839524]
7. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature 2021 Apr;592(7855):629-633. [doi: 10.1038/s41586-021-03430-5] [Medline: 33828294]
8. Azizi Z, Lindner S, Shiba Y, et al. A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health. Sci Rep 2023 Jul 17;13(1):11540. [doi: 10.1038/s41598-023-38457-3] [Medline: 37460705]
9. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One 2011;6(12):e28071. [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
10. Kaur D, Sobiesk M, Patil S, et al. Application of Bayesian networks to generate synthetic health data. J Am Med Inform Assoc 2021 Mar 18;28(4):801-811. [doi: 10.1093/jamia/ocaa303] [Medline: 33367620]
11. Mavrogenis AF, Scarlat MM. Artificial intelligence publications: synthetic data, patients, and papers. Int Orthop 2023 Jun;47(6):1395-1396. [doi: 10.1007/s00264-023-05830-w] [Medline: 37162553]
12. Meeker D, Kallem C, Heras Y, Garcia S, Thompson C. Case report: evaluation of an open-source synthetic data platform for simulation studies. JAMIA Open 2022 Oct;5(3):ac067. [doi: 10.1093/jamiaopen/ooac067] [Medline: 35958672]
13. Brownstein JS, Chu S, Marathe A, et al. Combining participatory influenza surveillance with modeling and forecasting: three alternative approaches. JMIR Public Health Surveill 2017 Nov 1;3(4):e83. [doi: 10.2196/publichealth.7344] [Medline: 29092812]
14. Guillaudeux M, Rousseau O, Petot J, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. NPJ Digit Med 2023 Mar 10;6(1):37. [doi: 10.1038/s41746-023-00771-5] [Medline: 36899082]
15. El Emam K. Status of synthetic data generation for structured health data. JCO Clin Cancer Inform 2023 Jun;7:e2300071. [doi: 10.1200/CCI.23.00071] [Medline: 37390378]
16. D'Amico S, Dall'Olio D, Sala C, et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. JCO Clin Cancer Inform 2023 Jun;7:e2300021. [doi: 10.1200/CCI.23.00021] [Medline: 37390377]
17. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. PLOS Digit Health 2023 Jan;2(1):e0000082. [doi: 10.1371/journal.pdig.0000082] [Medline: 36812604]

18. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med 2023 Oct 9;6(1):186. [doi: 10.1038/s41746-023-00927-3] [Medline: 37813960]
19. Ursin G, Sen S, Mottu JM, Nygård M. Protecting privacy in large datasets—first we assess the risk; then we fuzzy the data. Cancer Epidemiol Biomarkers Prev 2017 Aug 1;26(8):1219-1224. [doi: 10.1158/1055-9965.EPI-17-0172] [Medline: 28754793]
20. Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of supervised machine learning using synthetic data in health care: model to preserve privacy for data sharing. JMIR Med Inform 2020 Jul 20;8(7):e18910. [doi: 10.2196/18910] [Medline: 32501278]
21. Summers C, Griffiths F, Cave J, Panesar A. Understanding the security and privacy concerns about the use of identifiable health data in the context of the COVID-19 pandemic: survey study of public attitudes toward COVID-19 and data-sharing. JMIR Form Res 2022 Jul 7;6(7):e29337. [doi: 10.2196/29337] [Medline: 35609306]
22. 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. [doi: 10.1136/bmjopen-2020-043497] [Medline: 33863713]
23. Krenmayr L, Frank R, Drobig C, et al. GANerAid: realistic synthetic patient data for clinical trials. Inform Med Unlocked 2022;35:101118. [doi: 10.1016/j.imu.2022.101118]
24. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med 2020 Nov 9;3(1):147. [doi: 10.1038/s41746-020-00353-9] [Medline: 33299100]
25. Santos M. How to generate real-world synthetic data with CTGAN. Medium. 2023. URL: https://medium.com/towards-data-science/how-to-generate-real-world-synthetic-data-with-ctgan-af41b4d60fde [accessed 2024-06-04]
26. Ben-Aharon O, Magnezi R, Leshno M, Goldstein DA. Median survival or mean survival: which measure is the most appropriate for patients, physicians, and policymakers? Oncologist 2019 Nov;24(11):1469-1478. [doi: 10.1634/theoncologist.2019-0175] [Medline: 31320502]
27. Smith A, Lambert PC, Rutherford MJ. Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. BMC Med Res Methodol 2022 Jun 23;22(1):176. [doi: 10.1186/s12874-022-01654-1] [Medline: 35739465]
28. Breiman L, editor. Classification and Regression Trees: Chapman and Hall; 1998.
29. Breiman L. Random forests. Mach Learn 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
30. Pearl J. Bayesian networks: a model of self-activated memory for evidential reasoning. Presented at: Proceedings of the 7th Conference of the Cognitive Science Society; Aug 15 to 17, 1985; Irvine, CA URL: https://ftp.cs.ucla.edu/tech-report/198_-reports/850017.pdf [accessed 2024-06-04]
31. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. arXiv. Preprint posted online on Jul 1, 2019. [doi: 10.48550/arXiv.1907.00503]
32. Hayes T, Usami S, Jacobucci R, McArdle JJ. Using classification and regression trees (CART) and random forests to analyze attrition: results from two simulations. Psychol Aging 2015 Dec;30(4):911-929. [doi: 10.1037/pag0000046] [Medline: 26389526]
33. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. arXiv. Preprint posted online on Jun 10, 2016 URL: http://arxiv.org/abs/1606.03498 [accessed 2024-06-04] [doi: 10.48550/arXiv.1606.03498]
34. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol 2020 May 7;20(1):108. [doi: 10.1186/s12874-020-00977-1] [Medline: 32381039]
35. El Emam K, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data: O'Reilly Media; 2020.

Abbreviations

BN: Bayesian network
CART: classification and regression trees
CTGAN: conditional tabular generative adversarial network
GAN: generative adversarial network
HR: hazard ratio
HRD: hazard ratio distance
KM: Kaplan Meier
MST: median survival time
MSTA: median survival time of actual data
MSTS: median survival time of synthetic data
OS: overall survival
PFD: progression-free survival
RF: random forest
RWD: real-world data

SPD: synthetic patient data

Edited by C Lovis; submitted 03.12.23; peer-reviewed by D Hu, J Song; revised version received 06.04.24; accepted 08.05.24; published 18.06.24.

Please cite as:
Akiya I, Ishihara T, Yamamoto K
Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study
JMIR Med Inform 2024;12:e55118
URL: https://medinform.jmir.org/2024/1/e55118
doi:10.2196/55118

The Effect of an Electronic Medical Record–Based Clinical Decision Support System on Adherence to Clinical Protocols in Inflammatory Bowel Disease Care: Interrupted Time Series Study

Reed Taylor Sutton[1], MSc; Kaitlyn Delaney Chappell[1], MSc; David Pincock[2], MBA, MScIB; Daniel Sadowski[1], MD; Daniel C Baumgart[1], MBA, MD, PhD; Karen Ivy Kroeker[1], MSc, MD

1

2

Corresponding Author:
Karen Ivy Kroeker, MSc, MD

Abstract

Background: Clinical decision support systems (CDSSs) embedded in electronic medical records (EMRs), also called electronic health records, have the potential to improve the adoption of clinical guidelines. The University of Alberta Inflammatory Bowel Disease (IBD) Group developed a CDSS for patients with IBD who might be experiencing disease flare and deployed it within a clinical information system in 2 continuous time periods.

Objective: This study aims to evaluate the impact of the IBD CDSS on the adherence of health care providers (ie, physicians and nurses) to institutionally agreed clinical management protocols.

Methods: A 2-period interrupted time series (ITS) design, comparing adherence to a clinical flare management protocol during outpatient visits before and after the CDSS implementation, was used. Each interruption was initiated with user training and a memo with instructions for use. A group of 7 physicians, 1 nurse practitioner, and 4 nurses were invited to use the CDSS. In total, 31,726 flare encounters were extracted from the clinical information system database, and 9217 of them were manually screened for inclusion. Each data point in the ITS analysis corresponded to 1 month of individual patient encounters, with a total of 18 months of data (9 before and 9 after interruption) for each period. The study was designed in accordance with the Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI) guidelines for health informatics evaluations.

Results: Following manual screening, 623 flare encounters were confirmed and designated for ITS analysis. The CDSS was activated in 198 of 623 encounters, most commonly in cases where the primary visit reason was a suspected IBD flare. In Implementation Period 1, before-and-after analysis demonstrates an increase in documentation of clinical scores from 3.5% to 24.1% (P<.001), with a statistically significant level change in ITS analysis (P=.03). In Implementation Period 2, the before-and-after analysis showed further increases in the ordering of acute disease flare lab tests (47.6% to 65.8%; P<.001), including the biomarker fecal calprotectin (27.9% to 37.3%; P=.03) and stool culture testing (54.6% to 66.9%; P=.005); the latter is a test used to distinguish a flare from an infectious disease. There were no significant slope or level changes in ITS analyses in Implementation Period 2. The overall provider adoption rate was moderate at approximately 25%, with greater adoption by nurse providers (used in 30.5% of flare encounters) compared to physicians (used in 6.7% of flare encounters).

Conclusions: This is one of the first studies to investigate the implementation of a CDSS for IBD, designed with a leading EMR software (Epic Systems), providing initial evidence of an improvement over routine care. Several areas for future research were identified, notably the effect of CDSSs on outcomes and how to design a CDSS with greater utility for physicians. CDSSs for IBD should also be evaluated on a larger scale; this can be facilitated by regional and national centralized EMR systems.

(JMIR Med Inform 2024;12:e55314) doi:10.2196/55314

KEYWORDS

decision support system; clinical; electronic medical records; electronic health records; health record; medical record; EHR; EHRs; EMR; EMRs; decision support; CDSS; internal medicine; gastroenterology; gastrointestinal; implementation science; implementation; time series; interrupted time series analysis; inflammatory bowel disease; IBD; bowel; adherence; flare; flares; steroid; steroids; standardized care; nurse; clinical practice guidelines; chart; electronic chart; electronic medical chart

Introduction

Limited or delayed adoption of professional society–developed clinical care guidelines into practice is a common problem in medicine [1,2]. In 2007, researchers estimated that it took 17 years on average for only 14% of published evidence in guidelines to be translated into clinical practice [3,4]. One purported reason is that clinical guidelines by themselves are not actionable, as they largely describe what to do but not how to do it [5,6].

Clinical decision support systems (CDSSs) are tools that can be used to support provider decision-making. A CDSS uses clinical, patient, and other health information to supply providers with recommendations to assist in a variety of aspects of care, including diagnosis, treatment, and management [7,8]. Recent systematic reviews suggest that the use of CDSSs in clinical settings can improve practitioner performance in relation to adherence to best practice guidelines [7,9].

There are several demonstrated gaps in the adoption of professional society clinical care guidelines and best practices for inflammatory bowel disease (IBD). These include practices in medication management, preventative care, and bone health [10,11]. The University of Alberta IBD outpatient clinic (Edmonton, Alberta, Canada) has previously developed and implemented several clinical care pathways to consolidate best practices for IBD [10,12]. To further increase adoption, a clinical decision support (CDS) project was undertaken to integrate the pathways into the local electronic medical record (EMR). There are thousands of CDS projects built and deployed within commercial EMRs [13,14], yet there are few published evaluations of EMR-based CDSSsfor IBD [15,16]. Consequently, the objective of this pilot study was to evaluate the effectiveness and provider acceptance of an EMR-integrated CDSS in the context of IBD.

Methods

Ethical Considerations

This study received approval from the University of Alberta Health Research Ethics Board (Pro00083538). A waiver of informed consent was also approved as part of our study by the Health Research Ethics Board.

Organizational Setting

The study was conducted in the Comprehensive Academic Outpatient Center at the University of Alberta Hospital, which provides care for patients with IBD in the Greater Edmonton region as well as rural and remote communities across Alberta, Canada. It also serves a small number of patients with IBD from Saskatchewan, Northwest Territories, and British Columbia.

System Details and System in Use

The clinic's preexisting system was an enterprise EMR based on the 2014 version of Epic EMR (Epic Systems), which was being used for outpatient medical care in Edmonton, Alberta. This system was customized and branded locally as eCLINICIAN. Medication lists, allergies, and health problems are recorded and shared between users as part of clinical documentation, order entry, and planning. The system was implemented for gastroenterology outpatient care in March 2014.

As Epic is a general-purpose EMR, it includes built-in CDS functionality. For example,this includes generic functionality, such as alerting users when duplicate orders exist. More specialty-specific CDS features are often customized at the request and guidance of end users.

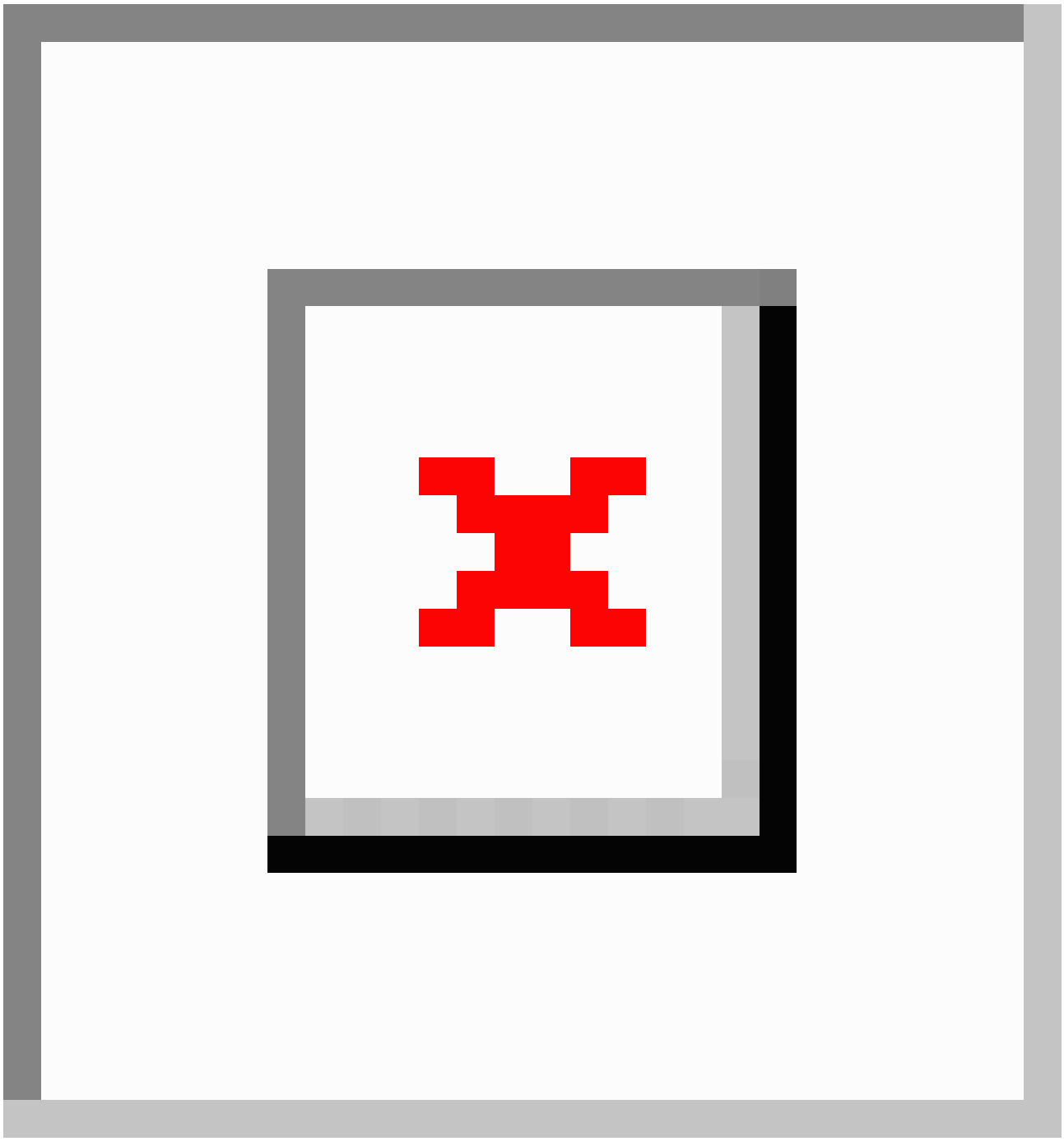
Functionality can be administered through a number of tools, including those referred to by Epic as "Flowsheets" (documentation tables), "Best Practice Advisories (BPAs)" (alerts) [17], and "SmartSets" (ie, grouping of orders and clinical content) [18].

These tools, particularly BPAs and SmartSets, are clinical data and test result driven; they can be triggered by unique combinations of provider characteristics, patient demographics, test results, clinical problems, as well as current and requested medications.

System Interruption and Intervention

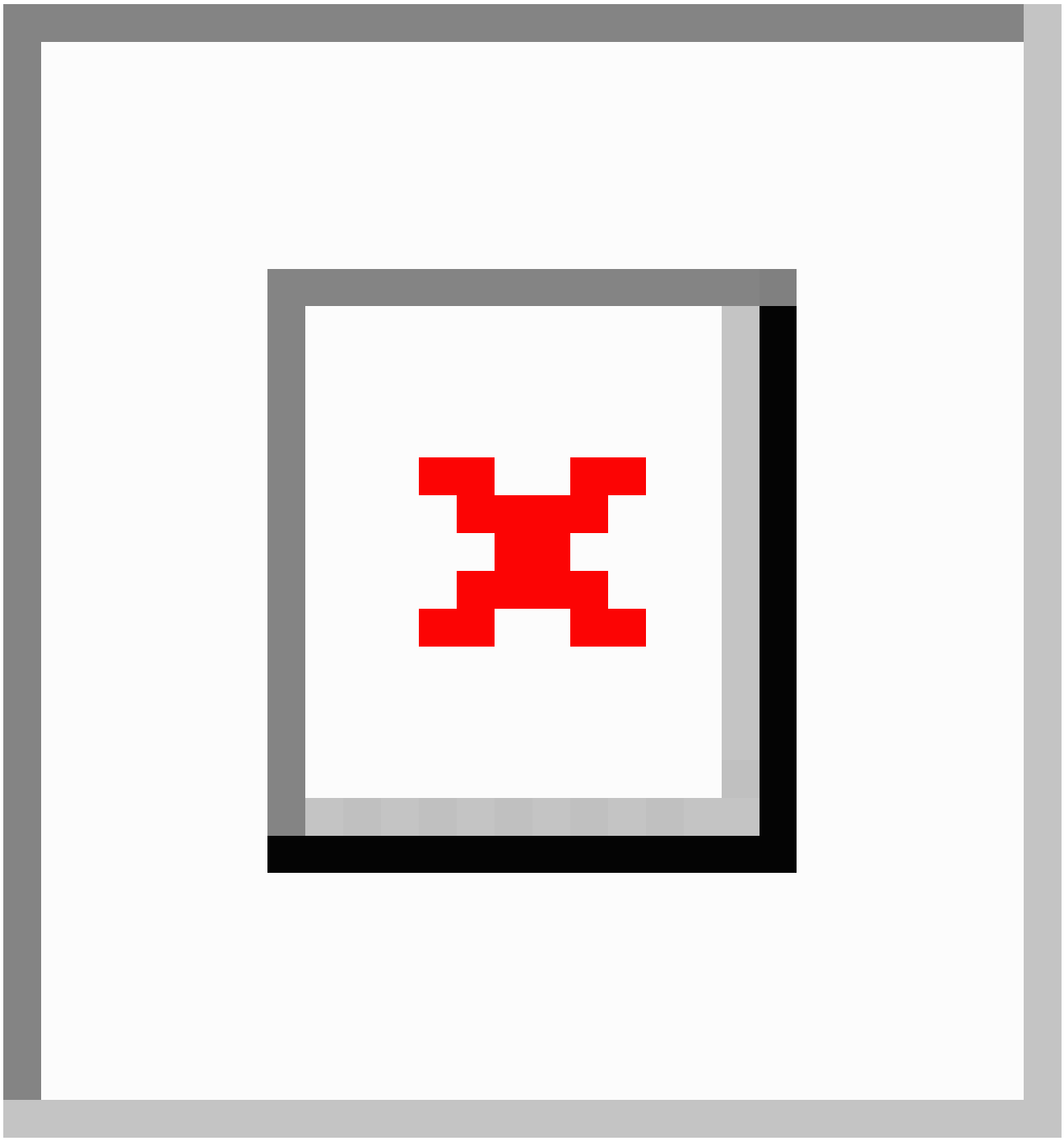
The system interruption and intervention uses BPA appearing in the clinician's navigator workflow. The BPA is triggered by the existence of IBD in the patient problem list or visit diagnosis fields. The BPA (Figure 1) prompts the clinician to complete clinical symptom indices—modified Harvey Bradshaw Index (mHBI) [19] for Crohn disease or partial Mayo (pMayo) score [20] for ulcerative colitis—for the encounter. If the score is indicative of a disease flare, the BPA instructs the user to activate a corresponding SmartSet.

Figure 1. Snapshot of the inflammatory bowel disease (IBD) flare clinical decision support system, showing the initial Best Practice Advisory. Best Practice Advisories act as alerts that present targeted patient-specific guidance to users. They can be active (disruptive pop-ups) or passive (navigation workflow) and can link to actions such as placing orders, order sets, initiating a care plan, or sending a message. This alert appeared passively in the providers' workflow navigation whenever IBD was in the patient problem list.



The SmartSet offers ordering and printing of appropriate lab panels, stool cultures, and other investigations, including imaging, procedures, and medication prescriptions. All recommendations were designed to be consistent with established IBD care guidelines and the flare protocol for the clinic. For example, during a flare encounter, the IBD flare lab panel and fecal calprotectin (FCP) tests are automatically selected for ordering (they can still be deselected by the provider). A snapshot of the SmartSet portion of the CDSS is shown in Figure 2.

Figure 2. Snapshot of the inflammatory bowel disease (IBD) flare clinical decision support system, showing the SmartSet, after activation by Best Practice Advisory. Not all sections of the SmartSet are shown, including sections for medications, imaging investigations, billing, and follow-up appointment booking. ALT: alanine transaminase; AST: aspartate aminotransferase; Cl: chloride; CO2: carbon dioxide; ESR: erythrocyte sedimentation rate; K: potassium; Na: sodium; NO DIFF: no differential.



Study Design

The study used a pre- and postimplementation interrupted time series (ITS) design, the interruption being the enhanced CDSS used within the EMR. Each data point represented 1 month of clinical encounters. For each intervention period, there was a total of 18 data points, 9 before and 9 after the intervention. Multimedia Appendix 1 presents an elaboration on the rationale for using an ITS design.

Physicians at the participating clinic were not guaranteed to have outpatient clinics on a weekly basis due to their service rotation; therefore, it was decided to aggregate the data points by month instead of by week. This avoided the potential week-to-week variation and ensured an adequate number of individual patient encounters (IBD flares) for each data point.

The Quality Criteria for ITS Designs checklist was used in the study design and assessment of appropriateness [21], and the Statement on Reporting of Evaluation Studies in Health Informatics (STARE-HI) guidelines were used for health informatics evaluations [22,23].

Participants

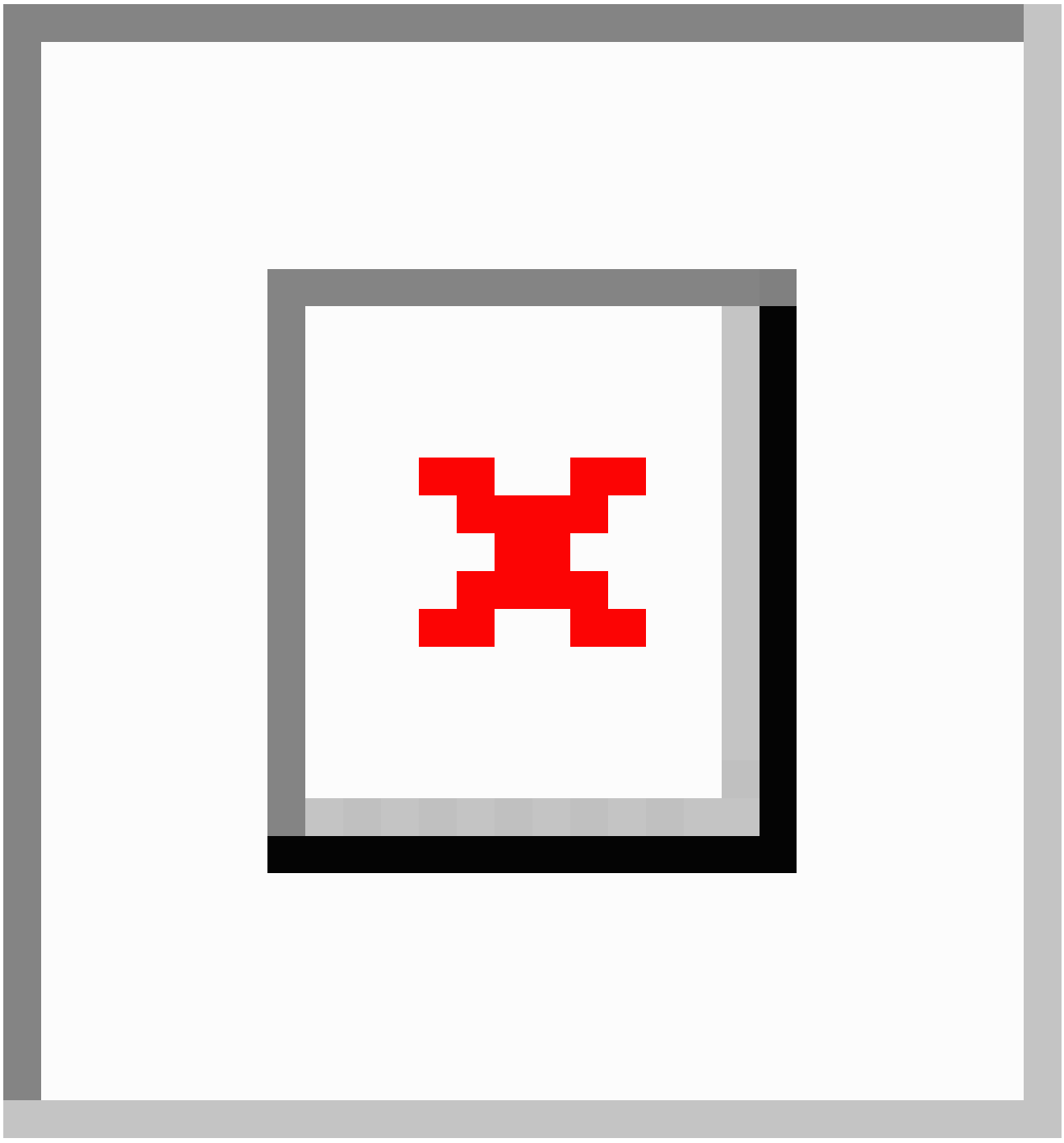
All IBD care providers at the university-based outpatient clinic were included in the study and invited to use the CDSS, including 7 IBD specialist clinicians, 1 IBD nurse practitioner, and 4 IBD specialist nurses. The term "IBD practitioner" will be used to collectively refer to IBD specialists and IBD nurse practitioners.

To be included in the data set, patients had to be under the care of the IBD providers; aged ≥18 years; and diagnosed with either Crohn disease or ulcerative colitis confirmed by imaging, pathology, or endoscopy report. They also had to be experiencing a flare of the disease during the included encounter, as defined by clinical scores (mHBI >5; pMayo >2) or noted symptoms in combination with physician judgment. Only initial encounters in a flare episode spanning multiple encounters were included.

Study Flow

The intervention was implemented and evaluated in 2 continuous periods (Figure 3). First, a pilot version was trialed by IBD nurses (Implementation Period 1), and then, the polished version was implemented across all providers in the division (ie, clinicians, nurse practitioners, and IBD nurses) as Implementation Period 2. The pilot version was trialed beginning in September 2017 and included the following 3 SmartSets available within the BPA, corresponding to different positions along the care path of a patient with flaring IBD: suspected flare, 2 to 4 weeks into the flare, and 16 weeks' postflare assessments. Feedback was gathered informally from providers (Multimedia Appendix 2) to inform further improvement to the CDSS.

Figure 3. Study design diagram of the 2-period interrupted time series design. First, the clinical decision support system (CDSS) was implemented as a limited pilot with inflammatory bowel disease (IBD) nurses (intervention 1), and then, it was fully implemented across all providers (intervention 2). Each data point (abbreviated as D) corresponds to 1 month of clinical encounters by study providers. NP: nurse practitioner.



After collecting feedback from the pilot, further changes were made to the CDSS. Aside from minor modifications to update included lab tests, the most significant change was the consolidation of the 3 separate SmartSets into 1, targeting the "suspected flare," the first step in the care pathway. The activation of the BPA in the initial CDSS was entirely manual and relied on the provider entering a specific visit diagnosis. However, in the full version, the BPA was set to automatically trigger based on the presence of an IBD diagnosis in the patient's problem list. This change was expected to improve the adoption and ease of use of the SmartSet for flare encounters.

The full implementation of the CDSS began on October 10, 2018. An instructional memo with paper-based workflow and educational material was sent to each provider (Multimedia Appendix 3). Over the course of 1 month, each participant was given the opportunity to ask questions about using the system and access to use the system in the sandbox environment. A demonstration of the system was also presented at weekly clinical rounds, with an opportunity to ask questions.

Outcome Measures

Process indicators were used to measure the proportion of adherent IBD practitioner flare encounters. These indicators include completion of clinical scores (mHBI or pMayo);

laboratory testing, such as standard lab panel, FCP, stool cultures, and *Clostridium difficile* toxin (only if diarrhea is present); and of vitamin D or calcium in conjunction with corticosteroid prescription, patient information given and documented, and modification of maintenance therapy. A secondary outcome was the adoption or acceptance of the CDSS measured by application rate (ratio of CDSS uses to CDSS available for activation).

Methods for Data Acquisition and Measurement

Potential encounters in the pre- and postintervention periods were initially identified by querying the eCLINICIAN EMR database for encounters with the included IBD providers, where patients had documentation of IBD in their problem list or diagnosis field (*International Classification of Diseases* coding). A sampling method was used to exclude encounters with specific reasons for visit deemed unlikely to constitute a flare based on exploratory analysis of the data set. Examples of excluded reasons for the visit included "medication refill," "medical insurance coverage," and "review results" (a more detailed description of the sampling method is available in a previous publication [10]). Encounters were then screened manually for inclusion and exclusion eligibility by one of the authors (RTS) and a research assistant.

Data for primary outcome measures were also queried and extracted from the EMR database, in collaboration with the eCLINICIAN reporting team in Alberta Health Services (AHS). The various database codes and IDs as well as the final SQL queries used to extract data are included in the Multimedia Appendix 4.

Methods for Data Analysis

Descriptive statistics were calculated to determine patient characteristics, with data presented as counts and proportions for categorical variables, mean (SD) values for normally distributed continuous variables, and median (IQR) values for nonnormally distributed continuous variables. Proportions were compared by using the Pearson χ^2 test [24].

A segmented regression analysis was performed for each primary outcome variable to determine the level and slope in the preintervention period as well as the change in level and slope in the postintervention period regarding the mean percentage of adherent encounters [25]. Autocorrelation in the residuals was tested using the Durbin-Watson test.

Data analysis was performed using IBM SPSS Statistics (version 23; IBM Corp) and R 3.5.1 (RStudio Inc) [26]. A 95% CI was used in all analyses unless otherwise specified.

Sample Size Determination

The sample size was first calculated for pre- and postimplementation cohorts based on logistic regression (Multimedia Appendix 5). With a power of 0.80 and a type I error set to 5%, the sample size required was approximately 634 for small effects and 145 for medium effects [27]. This assumes equal sample sizes (N) in the comparison groups and an initial proportion of adherence to each guideline component of approximately 70%, chosen based on a recent study by Jackson et al [11]. The sample size was calculated using G*Power 3.2.9.2 [28].

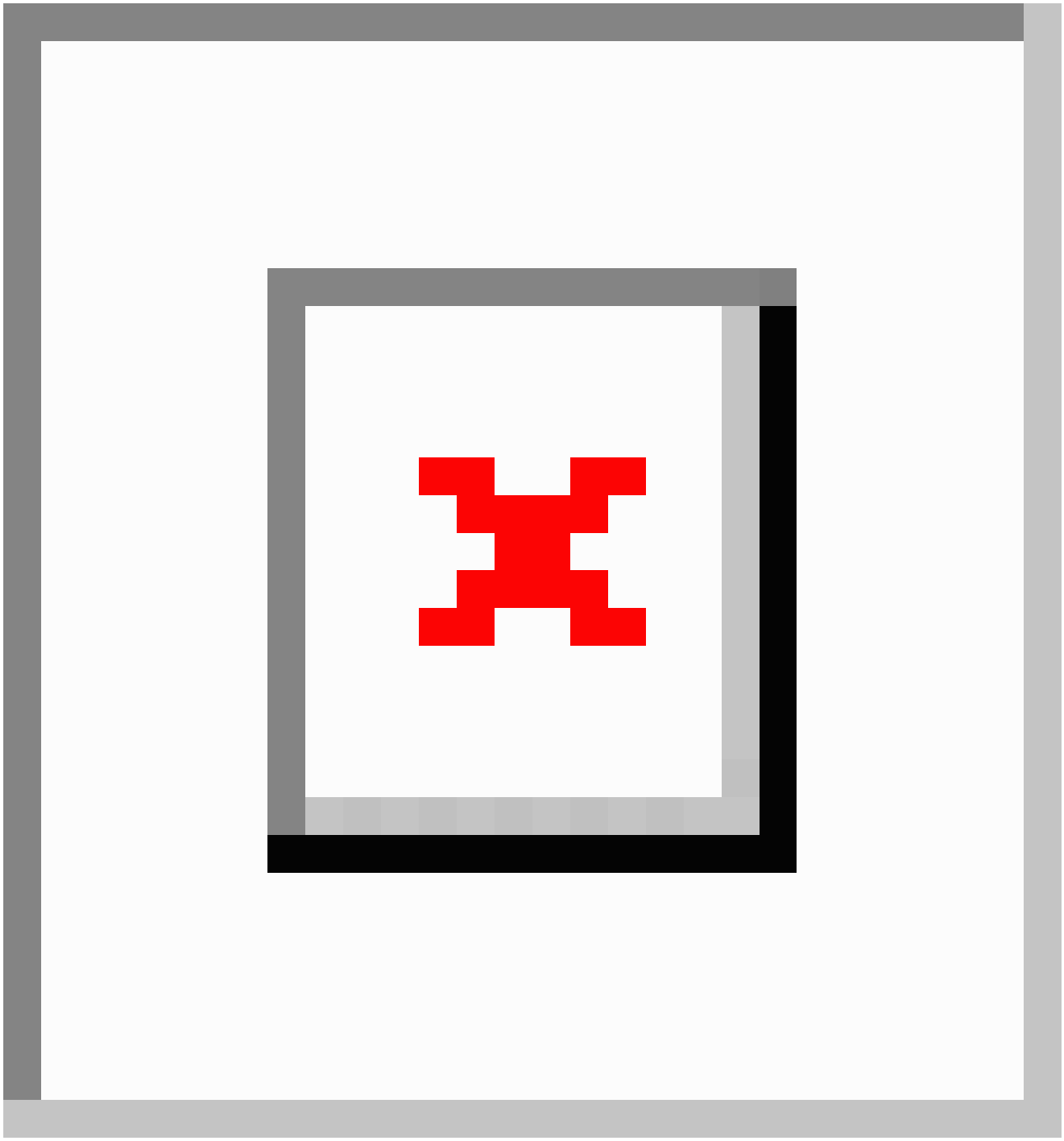
There is no standard method for determining power in time series analyses. However, a simulation-based power calculation displayed that with N=16 (8 data points in the preintervention period and 8 data points in the postintervention period), there is a 70% chance to detect an effect size of 0.5 or more, and over 90% chance to detect an effect size of 1 or more, at an alpha level of .05 [29]. It is also generally recommended in the literature to have over 100 observations per data point [25,30].

Results

Initial Data Set and Preprocessing

Figure 4 shows the study's flow diagram. The complete, extracted data set includes 31,726 encounters from January 1, 2017, to June 30, 2019. When considering only clinic visits (7655), orders (16,485), and telephone (5220) encounter types, the data set totals 29,360 (92.5%) encounters. There was an average of 998 encounters per month, with a minimum of 735 (December 2018) and a maximum of 1202 (May 2017) encounters. Of note, there is an overlap between both implementation periods (Figure 3), and thereby, a number of flare encounters appear in both analyses.

Figure 4. Flow data diagram for data extraction, screening, and analyses. CDSS: clinical decision support system.



Demographics of CDSS-Enabled Encounters

From September 2017 to June 2019, the CDSS was activated a total of 214 times across 214 encounters with 207 patients. Of these, 16 encounters were excluded from analysis due to, upon review, not being used appropriately for a flare or suspected flare encounter with a patient with IBD. This left 198 encounters, which are detailed in Table 1. More detailed demographics of providers using the system are included in Multimedia Appendix 6.

Table . Demographics of users and encounters invoking the inflammatory bowel disease (IBD) flare clinical decision support system.

Demographic variables	Study population (n=198)
Provider characteristics	
Provider type, n (%)	
IBD nurse	172 (86.9)
IBD practitioner	26 (13.1)
Patient characteristics	
Sex, n (%)	
Female	113 (57.1)
Male	85 (42.9)
Age (years), median (IQR)	37.5 (29-49)
Current IBD therapy, n (%)	
None	37 (18.7)
5-aminosalicylic acid only	53 (26.8)
Immunomodulator	18 (9.1)
Biologic monotherapy	59 (29.8)
Biologic combination therapy	31 (15.7)
Encounter characteristics	
Encounter type, n (%)	
Telephone	139 (70.2)
Orders only	32 (16.2)
Clinic visit	27 (13.6)
First encounter diagnosis, n (%)	
None	172 (86.9)
Crohn disease	11 (5.6)
Ulcerative colitis	10 (5.1)
Bloody diarrhea	2 (1.0)
IBD	1 (0.5)
Abdominal bloating	1 (0.5)
Ankylosing spondylitis	1 (0.5)
Visit reason, n (%)	
Suspected IBD flare	113 (57.1)
IBD	39 (19.7)
Disease flare-up	15 (7.6)
None	9 (4.5)
Referral	9 (4.5)
Follow-up	7 (3.5)
Diarrhea	3 (1.5)
Medication change	1 (0.5)
Medication problem	1 (0.5)

Study Findings and Outcome Data

Exploratory Analysis of Adherence to Clinical Protocols

Symptom Documentation

Of 192 patients with clinical scores (mHBI or pMayo) that were applicable (excluding those without pouch or short bowel or those newly diagnosed), 133 (69.3%) had a clinical score completed and documented in their chart at the index dispensation. Of all 198 encounters, 196 (99.0%) had symptoms (ie, pain, number and characteristics of stool, and the presence of blood) documented in the chart by the provider.

Laboratory Investigations

Full flare lab panels, including complete blood count, ferritin, electrolytes, creatinine, albumin, alkaline phosphatase, alanine transaminase, aspartate transaminase, and C-reactive protein (CRP), were ordered for 109/198 (55.1%) patients exactly at the encounter. Including orders up to 1 month prior, full panels were ordered for 183/198 (92.4%) patients. However, 113/198 (57.1%) had at least a partial lab panel, including complete blood count and CRP, ordered at the encounter, and 193/198 (97.5%) had partial lab panels, including complete blood count and CRP ordered up to 1 month prior to the encounter.

FCP was ordered at the encounter for 147/198 (74.2%) patients and within 1 month of the encounter for a further 36/198 (18.2%). This leaves only 15 (7.6%) who had no evaluation of FCP at all. Furthermore, testing for *Clostridium difficile* infection was done in 164/198 (82.8%) patients and for stool cultures in 160/198 (80.8) patients. In 138 patients with liquid stool or diarrhea mentioned in the progress note, 127 (92%) had *Clostridium difficile* testing ordered and 123 (89.1%) had stool cultures ordered.

Provision of Steroid-Sparing Therapy and Osteoprotective Therapy

In this data set, only 12 (6.1%) patients were prescribed steroids at their encounter. Of these, 6 (50%) had maintenance IBD therapy adjusted or added. In contrast, 37 (20%) of the 185 patients who were not prescribed steroids had maintenance therapy adjusted (P=.02 for χ^2).

Vitamin D or calcium supplementation was recommended for 8/12 (67%) patients prescribed steroids and 8/10 (80%) when excluding patients with vitamin D or calcium supplementation already documented in their medication list.

Implementation Period 1: Pilot CDSS Version With IBD Nurses

Implementation Period 1 included data from January 2017 to June 2018 (18 months), where September 2017 and beyond were labelled as the active intervention months (postintervention). Of the total 623 confirmed flare encounters, 502 occurred during Implementation Period 1 (Figure 3). Table 2 compares outcome measures before and after the intervention using chi-square tests. Notably, there was a substantial increase in the proportion of flare encounters with completed clinical scores from 3.5% (8/228) to 24.1% (66/274) post intervention. There was also an increase in the proportion of flare encounters with FCP ordered, from 16.7% (38/228) to 27% (74/274).

Table . Before-and-after analysis of process measures from Implementation Period 1.

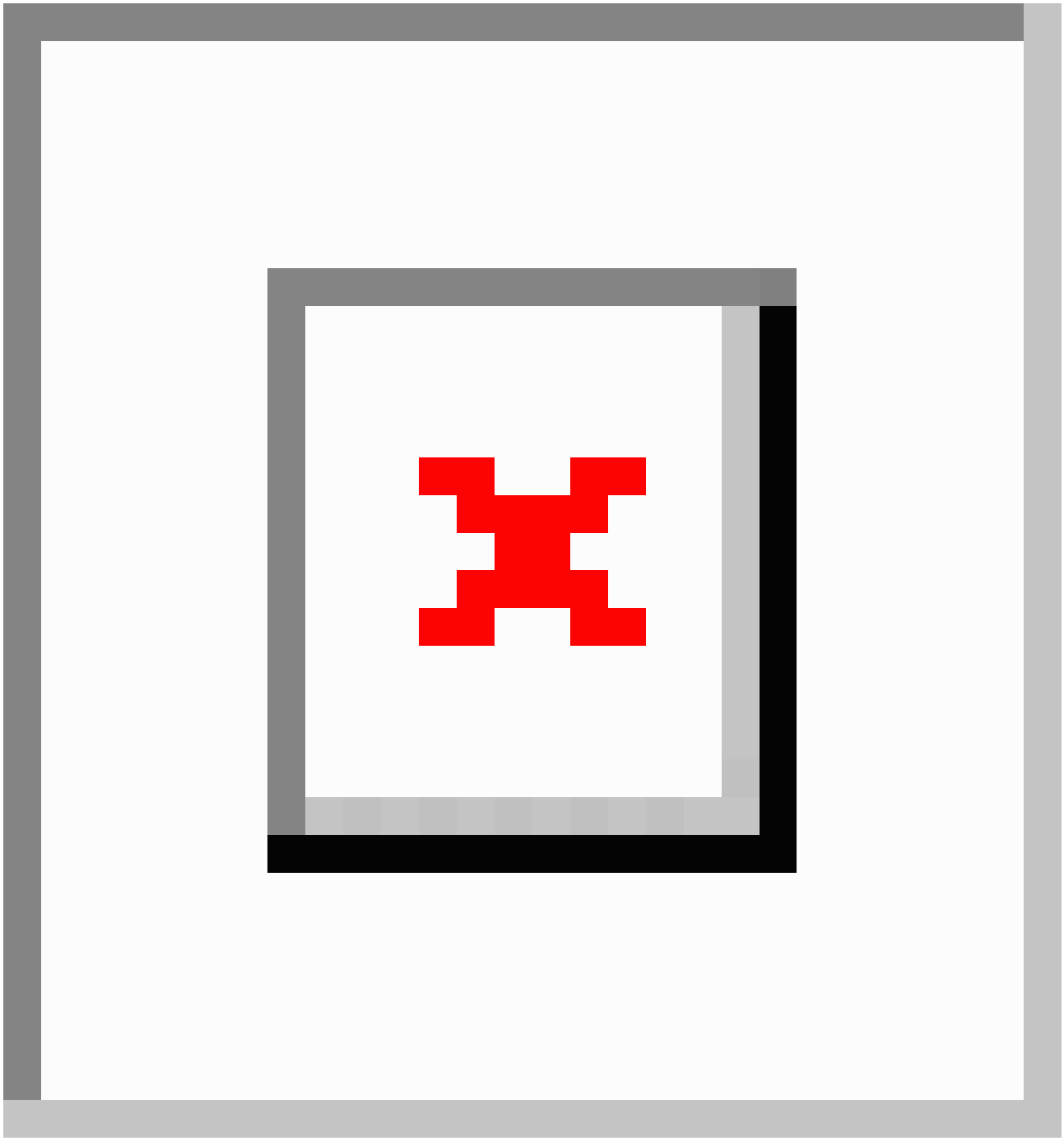
Parameter	Preintervention (n=228), n (%)	Postintervention (n=274), n (%)	*P* value[a]
CDSS[b] activated	0 (0)	66 (24.1)	<.001
Clinical score completed	8 (3.5)	66 (24.1)	<.001
Flare labs ordered	124 (54.4)	132 (48.2)	.33
C-reactive protein ordered	156 (68.4)	178 (65.0)	.56
Fecal calprotectin ordered	38 (16.7)	74 (27.0)	.048
Stool cultures ordered	128 (56.1)	162 (59.1)	.63
Clostridium difficile test ordered	128 (56.1)	172 (62.8)	.29

[a]*P* value of the Pearson chi-square test comparing proportions.

[b]CDSS: clinical decision support system.

ITS analysis was done for outcomes that were significant in the before-and-after analyses (Figure 5). For clinical score completion rates, there was no slope change (estimated β −1.22, 95% CI −4.44 to 2.01; P=.43), but there was a level increase (estimated β 19.0, 95% CI 2.39-35.60; P=.03). For calprotectin testing, there was no slope change (estimated β −2.45, 95% CI −6.21 to 1.32; P=.19) or level change (estimated β 14.77, 95% CI −4.63 to 34.17; P=.13).

Figure 5. Segmented regression for Implementation Period 1 (pilot) of the inflammatory bowel disease flare clinical decision support system on rates of (A) clinical score completion and (B) calprotectin testing.



Implementation Period 2: Full CDSS Implementation With All Providers

Implementation Period 2 included data from January 2018 to June 2019 (18 months), where October 2018 and beyond were postintervention months. Of the total 623 confirmed flare encounters, 492 occurred during Implementation Period 2 (Figure 3). Table 3 compares outcome measures before and after the intervention using chi-square tests. There were increases in the proportion of flare encounters with completed flare labs (109/229, 47.6% to 173/263, 65.8%), CRP ordered (147/229, 64.2% to 207/263, 78.7%), calprotectin ordered (64/229, 27.9% to 98/263, 37.3%), and stool cultures ordered (125/229, 54.6% to 176/263, 66.9%).

Table . Before-and-after analysis of process measures from Implementation Period 2.

Parameter	Preintervention (n=229), n (%)	Postintervention (n=263), n (%)	*P* value[a]
Application of SmartSets	52 (22.7)	72 (27.4)	.23
Clinical score completed	58 (25.3)	75 (28.5)	.43
Flare labs ordered	109 (47.6)	173 (65.8)	<.001
C-reactive protein ordered	147 (64.2)	207 (78.7)	<.001
Fecal calprotectin ordered	64 (27.9)	98 (37.3)	.03
Stool cultures ordered	125 (54.6)	176 (66.9)	.005
Clostridium testing ordered	136 (59.4)	177 (67.3)	.70

[a]*P* value of the Pearson chi-square test comparing proportions.

The ITS analysis for significant outcomes is shown in Figure 6, and accompanying β values for slope change and level change with 95% CIs are shown in Table 4. For Period 2, there were no slope or level increases that reached significance at *P*=.05, although CRP testing and stool culture testing would be significant for a level increase at *P*=.10.

Figure 6. Segmented regression for Implementation Period 2 of the inflammatory bowel disease (IBD) flare clinical decision support system on rates of (A) clinical score completion, (B) flare lab testing, (C) C-reactive protein testing, (D) calprotectin testing, (E) stool culture testing, and (F) *Clostridium difficile* testing.

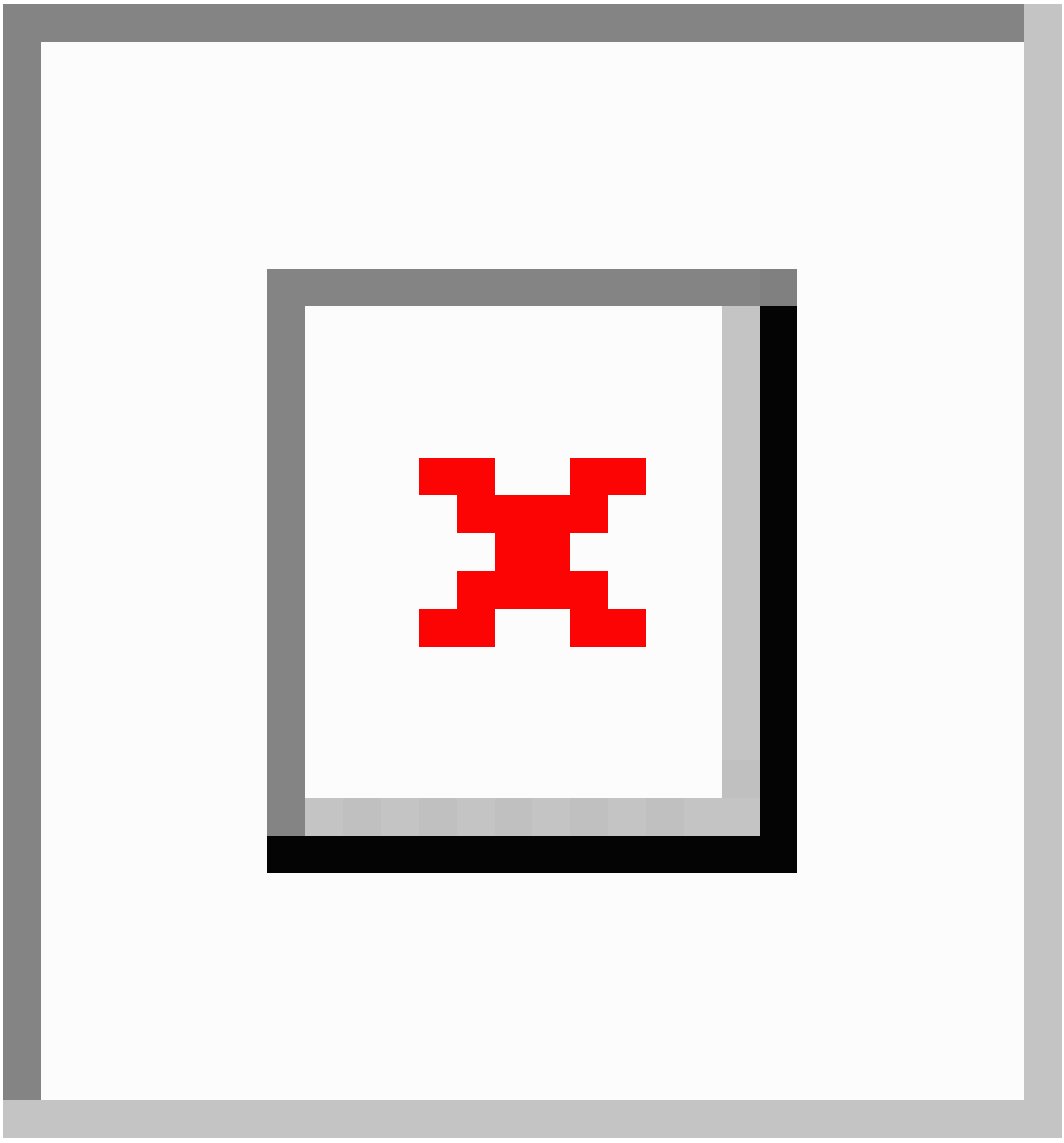


Table . Parameters for segmented logistic regression analysis of the inflammatory bowel disease (IBD) clinical decision support system (CDSS) in Implementation Period 2.

Parameter	ß	95% CI	*P* value
Application rate			
Preintervention slope (secular trend, per month)	0.151	−3.757 to 4.059	.94
Change in slope (gradual effect, per month)	2.019	−3.508 to 7.546	.45
Change in intercept (immediate effect)	−5.048	−33.86 to 23.76	.71
Clinical scores completed and documented			
Preintervention slope (secular trend, per month)	1.648	−1.596 to 4.893	.29
Change in slope (gradual effect, per month)	−2.463	−7.051 to 2.125	.27
Change in intercept (immediate effect)	−0.992	−24.91 to 22.92	.93
IBD flare lab tests ordered			
Preintervention slope (secular trend, per month)	−0.016	−2.693 to 2.662	.99
Change in slope (gradual effect, per month)	1.929	−1.858 to 5.715	.29
Change in intercept (immediate effect)	12.60	−7.137 to 32.34	.19
C-reactive protein ordered			
Preintervention slope (secular trend, per month)	−0.742	−3.121 to 1.637	.52
Change in slope (gradual effect, per month)	1.253	−2.111 to 4.618	.44
Change in intercept (immediate effect)	14.89	−2.645 to 32.43	.09
Fecal calprotectin ordered			
Preintervention slope (secular trend, per month)	1.298	−2.209 to 4.806	.44
Change in slope (gradual effect, per month)	0.183	−4.778 to 5.143	.94
Change in intercept (immediate effect)	−1.034	−26.89 to 24.82	.93
Stool cultures ordered			
Preintervention slope (secular trend, per month)	−1.060	−3.650 to 1.529	.40
Change in slope (gradual effect, per month)	1.714	−1.948 to 5.376	.33
Change in intercept (immediate effect)	15.37	−3.715 to 34.46	.11
Clostridium difficile ordered			
Preintervention slope (secular trend, per month)	−0.228	−2.613 to 2.158	.84
Change in slope (gradual effect, per month)	1.825	−1.549 to 5.198	.27
Change in intercept (immediate effect)	3.258	−14.33 to 20.84	.70

Discussion

Answering the Study Question

In this study, we evaluated the effectiveness of a CDSS that aimed to standardize protocols for patients with IBD experiencing an acute disease flare. An increase in several practices was demonstrated following the CDSS implementation, including increased FCP use. Completion of clinical scores also increased during Implementation Period 1 (before-and-after analysis and ITS analysis) and remained increased throughout Implementation Period 2.

We did not reach significance in slope changes or level changes in any ITS analysis in Period 2. This could be due to the sample size, which may also account for the large variance seen in some data points. There were, however, some encouraging upward trends in flare lab testing, particularly CRP (P<.10 in the ITS analysis) and stool cultures.

In characterizing the adoption of this CDSS by the application rate, an interesting finding was that the CDSS was used more by IBD nurses compared to nurse practitioners. This could represent the nurses' increased experience with the CDSS from the pilot phase and our CDSS focus on decisions related to patients experiencing a disease flare. In the University of Alberta clinic, patients are instructed to call the IBD nurse flare line if they experience changes in symptoms, and so nurses are often the first point of contact in the flare clinical pathway. This is supported by our data showing flare encounters are primarily telephone encounters. Other research has shown that flares are unlikely to coincide with scheduled clinic appointments, which aligns with the current uptake in remote monitoring and rapid access clinics [31-33].

Our observed CDSS use by specialized IBD nurses is in contrast to several other studies that have demonstrated that nurses are less likely to use CDSSs when making decisions about care they are experienced and confident in delivering, especially in the case of telephone triage decisions [34-36]. Our results could be a product of the integration of the nurses' feedback after the pilot phase, a strategy that may have increased the utility of the CDSS for nurses. This highlights recommendations from other research that emphasize the importance of engaging all stakeholders but especially end users in the CDSS design [37,38].

Limitations of the Study

There are several limitations to this research. Although the ITS design allows for better characterization of temporal changes compared to before-and-after analyses, it is still possible that other changes, such as clinic structure and release or dissemination of guidelines, could have led to the changes observed. However, apart from the intervention activation and the released memo and instructions for use that were disseminated, to our knowledge, there were no other educational campaigns, institutional changes, or major publications promoting the specific care guidelines investigated by the study. There were subtle changes in staff, for example, the joining of a new IBD physician and the leaving of another. However, there were no changes in IBD nurse staff, who were the primary users of the CDSS.

In contrast to the advantage of our 2-phased design regarding the opportunity for feedback from nurses, the design may have hindered our ability to demonstrate change. As we used the same group of IBD nurses in the pilot (Phase 1) and implementation (Phase 2) periods, our baseline use prior to the beginning of Phase 2 had already started. This may have accelerated the observed uptake speed of the CDSS by practitioners and could have also led to an underestimation of the changes before and after Implementation Phase 2.

Sample size is another limitation. In an ITS analysis, it is recommended to have a minimum of 16 data points and 100 observations per data point [25,29,30]. Although we met the data point requirement, the number of flares per month was consistently under 50. Future studies should aim to include more data points, which may require multisite participation. Unfortunately, at the time of this study, the EMR software was only deployed at a single site.

We only captured data from orders that were tied to the encounter. If a decision was made to not order labs for any reason (eg, they were recently completed), they would not be captured by our extraction. As a consequence, estimates of protocol adherence could be deflated.

Finally, it is important to note that for process measures that depend on manual data entry, such as clinical score completion, this research method can only determine whether a process was documented as completed but not necessarily whether it was actually completed. This may have resulted in underestimates of protocol adherence.

Future Directions

The currently available CDSS in this study was limited in its ability to support complex multiprovider pathways and tie together multiple visits along a pathway. More advanced CDSS workflows should be investigated in future versions of the CDSS software and evaluated for effectiveness.

Triggering logic for CDSSs should also be precisely targeted. For example, a CDSS should determine whether a patient has had a test done within a certain time span, and if not, prompt the user to order it. The reverse should also be possible; if a test has been recently ordered (eg, *Clostridium difficile*, which can only be tested once every 2 weeks), the CDSS could automatically deselect or prompt the user to remove this order to save downstream resources. This was not possible with the resources available in our CDSS environment.

In extracting data for analysis, a significant challenge was identifying flare encounters based on EMR data. The problem stems from a lack of discrete data identifying patients with active diseases (clinical scores were not regularly documented as discrete data). Future research should seek to develop a case definition for disease flare through administrative provincial data sets. This could include quantitative metrics, such as CRP and FCP, that predict the likelihood of flare, but it could also include the integration of a case-finding algorithm that uses natural language processing to parse clinical notes. This strategy

has been explored in several other diseases and has been shown to significantly improve case detection [39]. Some work has been done in IBD to identify phenotypic information from clinic notes using natural language processing [40].

The methodology used in this research should be expanded to investigate the effects of improved versions of CDSS for IBD on other community clinics and nonacademic practices throughout Alberta. Cluster-randomized designs or stepped-wedge designs could be explored since multiple clinics could be available for randomization.

This study did not investigate the impact on patient outcomes, which would require a longer follow-up period (ideally 2 or more years). Nonetheless, long-term patient outcomes for the CDSS are of great importance [9] and should be explored in the future.

Conclusions

Through our study, we designed and implemented, in 2 phases, a CDSS for IBD disease flare embedded in existing EMR software and evaluated the impact of the CDSS on provider adoption of clinical guidelines and local best practices. We have shown moderate adoption and acceptance of this system by providers, particularly by IBD nurses, as measured by the system application rate. Findings from the first phase support the hypothesis that the CDSS improved the use of FCP and the documentation of clinical scores. Findings from the second phase support further improvement in ordering flare lab panels, CRP, and stool cultures, as shown in before-and-after analysis and multivariate analysis. In addition, potential improvements in workflow integration were identified through qualitative questionnaires and feedback forms; areas for future research have also been established.

Acknowledgments

The authors acknowledge the faculty and staff of the inflammatory bowel disease (IBD) Unit and Division of Gastroenterology at the University of Alberta Hospital, who helped with the design and implementation of the IBD clinical care pathway (CCP). We also acknowledge the staff of Alberta Health Services (AHS) for their assistance with supplying the data. We thank Mr Darryl Wilson, the reporting systems analyst for the AHS Information Systems, for his assistance with the natural language queries (SQL) data acquisition from eCLINICIAN, and Mr Nathan Stern for helping with the chart review. Finally, we would like to thank the late Dr Richard Fedorak for his contribution to this work.

All results and inferences reported in this manuscript are independent of the funding and support sources.

Authors' Contributions

RTS contributed to study design, data collection, data analysis, and manuscript drafting. KDC contributed to drafting and revision of the manuscript. DP, DCS, and DCB contributed to the critical revision of the manuscript. KIK contributed to the study design as well as the analysis and critical revision of the manuscript. All authors approved the final version. KIK is the guarantor of the paper.

Conflicts of Interest

This study was supported by the Crohn’s and Colitis Canada via the Promoting Access and Care through Centres of Excellence (PACE) initiative. RTS was also supported by studentships from Alberta Innovates, the Faculty of Medicine and Dentistry, University of Alberta, and the Canadian Institutes of Health Research (CIHR). All other authors have no conflicts of interest to declare.

Multimedia Appendix 1
The rationale for using an interrupted time series design.
[DOCX File, 19 KB - medinform_v12i1e55314_app1.docx]

Multimedia Appendix 2
Provider feedback.
[DOCX File, 16 KB - medinform_v12i1e55314_app2.docx]

Multimedia Appendix 3
Materials distributed to providers.
[DOCX File, 781 KB - medinform_v12i1e55314_app3.docx]

Multimedia Appendix 4
eCLINICIAN query information.
[DOCX File, 68 KB - medinform_v12i1e55314_app4.docx]

Multimedia Appendix 5
Sample size calculation.
[DOCX File, 13 KB - medinform_v12i1e55314_app5.docx]

Multimedia Appendix 6
Demographics of users (inflammatory bowel disease nurses and practitioners).
[DOCX File, 13 KB - medinform_v12i1e55314_app6.docx]

References

1. Davis DA, Taylor-Vaisey A. Translating guidelines into practice. A systematic review of theoretic concepts, practical experience and research evidence in the adoption of clinical practice guidelines. CMAJ 1997 Aug 15;157(4):408-416. [Medline: 9275952]
2. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH. Why don't physicians follow a framework for improvement. J Am Med Assoc 1999 Oct;282:1458-1465. [doi: 10.1001/jama.282.15.1458] [Medline: 10535437]
3. Westfall JM, Mold J, Fagnan L. "Practice-based research - "blue highways" on the NIH roadmap". JAMA 2007 Jan 24;297(4):403-406. [doi: 10.1001/jama.297.4.403] [Medline: 17244837]
4. Balas EA, Boren SA. Managing clinical knowledge for health care improvement. Yearb Med Inform 2000(1):65-70. [Medline: 27699347]
5. Shortliffe T. Medical thinking: what should we do? Presented at: Conference on Medical Thinking; Jun 23, 2006; London, UK URL: https://slideplayer.com/slide/10838966/ [accessed 2024-03-05]
6. Vander Schaaf EB, Seashore CJ, Randolph GD. Translating clinical guidelines into practice: challenges and opportunities in a dynamic health care environment. N C Med J 2015 Sep;76:230-234. [doi: 10.18043/ncm.76.4.230] [Medline: 26509513]
7. Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
8. 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;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
9. Jaspers MWM, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. J Am Med Inform Assoc 2011 May 1;18(3):327-334. [doi: 10.1136/amiajnl-2011-000094] [Medline: 21422100]
10. Sutton RT. Adherence to clinical care protocols for inflammatory bowel disease and evaluation of a clinical decision support system to improve adherence [Master's thesis].: University of Alberta; 2019 URL: https://era.library.ualberta.ca/items/e4651c09-d19f-4d58-b7bc-eeb9368a974e [accessed 2024-03-05]
11. Jackson BD, Con D, Liew D, De Cruz P. Clinicians' adherence to international guidelines in the clinical care of adults with inflammatory bowel disease. Scand J Gastroenterol 2017 May;52(5):536-542. [doi: 10.1080/00365521.2017.1278785] [Medline: 28128675]
12. Lytvyak E, Sutton RT, Dieleman LA, Peerani F, Fedorak RN, Kroeker KI. Management of inflammatory bowel disease patients with clinical care pathways reduces emergency department utilization. Crohns Colitis 360 2020 Oct 13;2(4):taa080. [doi: 10.1093/crocol/otaa080] [Medline: 36777757]
13. Epic UserWeb. 2018. URL: https://comlib.epic.com [accessed 2018-05-08]
14. Pauwen NY, Louis E, Siegel C, Colombel JF, Macq J. Integrated care for Crohn's disease: a plea for the development of clinical decision support systems. J Crohns Colitis 2018 Nov 28;12(12):1499-1504. [doi: 10.1093/ecco-jcc/jjy128] [Medline: 30496446]
15. Breton J, Witmer CM, Zhang Y, et al. Utilization of an electronic medical record-integrated dashboard improves identification and treatment of anemia and iron deficiency in pediatric inflammatory bowel disease. Inflamm Bowel Dis 2021 Aug 19;27(9):1409-1417. [doi: 10.1093/ibd/izaa288] [Medline: 33165613]
16. Jackson B, Begun J, Gray K, et al. Clinical decision support improves quality of care in patients with ulcerative colitis. Aliment Pharmacol Ther 2019 Apr;49(8):1040-1051. [doi: 10.1111/apt.15209] [Medline: 30847962]
17. Epic Userweb. Best Practice Advisories Setup and Support Guide: Epic Userweb Software; 2018.
18. Epic Userweb. Decision Support Strategy Handbook: Epic Userweb Software; 2018.
19. Harvey RF, Bradshaw JM. A simple index of Crohn's disease activity. Lancet 1980 Mar 8;1(8167):514. [doi: 10.1016/s0140-6736(80)92767-1] [Medline: 6102236]
20. Lewis JD, Chuai S, Nessel L, Lichtenstein GR, Aberra FN, Ellenberg JH. Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. Inflamm Bowel Dis 2008 Dec;14(12):1660-1666. [doi: 10.1002/ibd.20520] [Medline: 18623174]

21. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. Int J Technol Assess Health Care 2003;19(4):613-623. [doi: 10.1017/s0266462303000576] [Medline: 15095767]
22. Talmon J, Ammenwerth E, Brender J, de Keizer N, Nykänen P, Rigby M. STARE-HI--Statement on reporting of evaluation studies in Health Informatics. Int J Med Inform 2009 Jan;78(1):1-9. [doi: 10.1016/j.ijmedinf.2008.09.002] [Medline: 18930696]
23. Brender J, Talmon J, de Keizer N, Nykänen P, Rigby M, Ammenwerth E. Statement on reporting of evaluation studies in Health Informatics. Appl Clin Inform 2013;04(3):331-358. [doi: 10.4338/ACI-2013-04-RA-0024]
24. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London Edinburgh Dublin Phil Mag J Sci 1900 Jul;50(302):157-175. [doi: 10.1080/14786440009463897]
25. 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]
26. Muggeo VMR. Segmented: an R package to fit regression models with broken-line relationships. R News 2008;7:1609-3631 [FREE Full text]
27. Chen H, Cohen P, Chen S. How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. Commun Stat - Simul C 2010 Mar 31;39(4):860-864. [doi: 10.1080/03610911003650383]
28. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and BIOMEDICAL sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
29. Zhang F, Wagner AK, Ross-Degnan D. Simulation-based power calculation for designing interrupted time series analyses of health policy interventions. J Clin Epidemiol 2011 Nov;64(11):1252-1261. [doi: 10.1016/j.jclinepi.2011.02.007] [Medline: 21640554]
30. Jandoc R, Burden AM, Mamdani M, Lévesque LE, Cadarette SM. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. J Clin Epidemiol 2015 Aug;68(8):950-956. [doi: 10.1016/j.jclinepi.2014.12.018] [Medline: 25890805]
31. Kemp K, Griffiths J, Campbell S, Lovell K. An exploration of the follow-up up needs of patients with inflammatory bowel disease. J Crohns Colitis 2013 Oct;7(9):e386-e395. [doi: 10.1016/j.crohns.2013.03.001] [Medline: 23541150]
32. Nene S, Gonczi L, Kurti Z, et al. Benefits of implementing a rapid access clinic in a high-volume inflammatory bowel disease center: access, resource utilization and outcomes. World J Gastroenterol 2020 Feb 21;26(7):759-769. [doi: 10.3748/wjg.v26.i7.759] [Medline: 32116423]
33. Pure N, Mize C. P193 the development of an IBD specialty clinic within a gastroenterology practice. Gastroenterology 2018 Jan;154(1):S107-S108. [doi: 10.1053/j.gastro.2017.11.253]
34. Dowding D, Mitchell N, Randell R, Foster R, Lattimer V, Thompson C. Nurses' use of computerised clinical decision support systems: a case site analysis. J Clin Nurs 2009 Apr;18(8):1159-1167. [doi: 10.1111/j.1365-2702.2008.02607.x] [Medline: 19320785]
35. O'Cathain A, Sampson FC, Munro JF, Thomas KJ, Nicholl JP. Nurses' views of using computerized decision support software in NHS direct. J Adv Nurs 2004 Feb;45(3):280-286. [doi: 10.1046/j.1365-2648.2003.02894.x] [Medline: 14720245]
36. O'Cathain A, Nicholl J, Sampson F, Walters S, McDonnell A, Munro J. Do different types of nurses give different triage decisions in NHS direct? A mixed methods study. J Health Serv Res Policy 2004 Oct;9(4):226-233. [doi: 10.1258/1355819042250221] [Medline: 15509408]
37. Rocque G, Miller-Sonnet E, Balch A, et al. Engaging multidisciplinary stakeholders to drive shared decision-making in oncology. J Palliat Care 2019 Jan;34(1):29-31. [doi: 10.1177/0825859718810723] [Medline: 30382006]
38. Daudelin DH, Ruthazer R, Kwong M, et al. Stakeholder engagement in methodological research: development of a clinical decision support tool. J Clin Transl Sci 2020 Apr;4(2):133-140. [doi: 10.1017/cts.2019.443] [Medline: 32313703]
39. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep;23(5):1007-1015. [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
40. South BR, Shen S, Jones M, et al. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics 2009 Sep 17;10 Suppl 9(Suppl 9):S12. [doi: 10.1186/1471-2105-10-S9-S12] [Medline: 19761566]

Abbreviations

AHS: Alberta Health Services
BPA: Best Practice Advisory
CDS: clinical decision support
CDSS: clinical decision support system
CRP: C-reactive protein

EMR: electronic medical record
FCP: fecal calprotectin
IBD: inflammatory bowel disease
ITS: interrupted time series
mHBI: modified Harvey Bradshaw Index
pMayo: partial Mayo
STARE-HI: Statement on Reporting of Evaluation Studies in Health Informatics

Edited by C Lovis; submitted 13.12.23; peer-reviewed by T Xenodemetropoulos; accepted 02.02.24; published 22.03.24.

Please cite as:
Sutton RT, Chappell KD, Pincock D, Sadowski D, Baumgart DC, Kroeker KI
The Effect of an Electronic Medical Record–Based Clinical Decision Support System on Adherence to Clinical Protocols in Inflammatory Bowel Disease Care: Interrupted Time Series Study
JMIR Med Inform 2024;12:e55314
URL: https://medinform.jmir.org/2024/1/e55314
doi:10.2196/55314

Implementable Prediction of Pressure Injuries in Hospitalized Adults: Model Development and Validation

Thomas J Reese[1], PharmD, PhD; Henry J Domenico[2], MS; Antonio Hernandez[3], MD, MSCI; Daniel W Byrne[1,2], MS; Ryan P Moore[2], MS; Jessica B Williams[4], RN, BSN, CCRN, APRN; Brian J Douthit[1], RN-BC, PhD; Elise Russo[1], MPH, PMP; Allison B McCoy[1], PhD; Catherine H Ivory[1], PhD, RN; Bryan D Steitz[1], PhD; Adam Wright[1], PhD

1

2

3

4

Corresponding Author:
Thomas J Reese, PharmD, PhD

Abstract

Background: Numerous pressure injury prediction models have been developed using electronic health record data, yet hospital-acquired pressure injuries (HAPIs) are increasing, which demonstrates the critical challenge of implementing these models in routine care.

Objective: To help bridge the gap between development and implementation, we sought to create a model that was feasible, broadly applicable, dynamic, actionable, and rigorously validated and then compare its performance to usual care (ie, the Braden scale).

Methods: We extracted electronic health record data from 197,991 adult hospital admissions with 51 candidate features. For risk prediction and feature selection, we used logistic regression with a least absolute shrinkage and selection operator (LASSO) approach. To compare the model with usual care, we used the area under the receiver operating curve (AUC), Brier score, slope, intercept, and integrated calibration index. The model was validated using a temporally staggered cohort.

Results: A total of 5458 HAPIs were identified between January 2018 and July 2022. We determined 22 features were necessary to achieve a parsimonious and highly accurate model. The top 5 features included tracheostomy, edema, central line, first albumin measure, and age. Our model achieved higher discrimination than the Braden scale (AUC 0.897, 95% CI 0.893-0.901 vs AUC 0.798, 95% CI 0.791-0.803).

Conclusions: We developed and validated an accurate prediction model for HAPIs that surpassed the standard-of-care risk assessment and fulfilled necessary elements for implementation. Future work includes a pragmatic randomized trial to assess whether our model improves patient outcomes.

(JMIR Med Inform 2024;12:e51842) doi:10.2196/51842

KEYWORDS

patient safety; electronic health record; EHR; implementation; predictive analytics; prediction; injury; pressure injury; hospitalization; adult; development; routine care; prediction model; pressure sore

Introduction

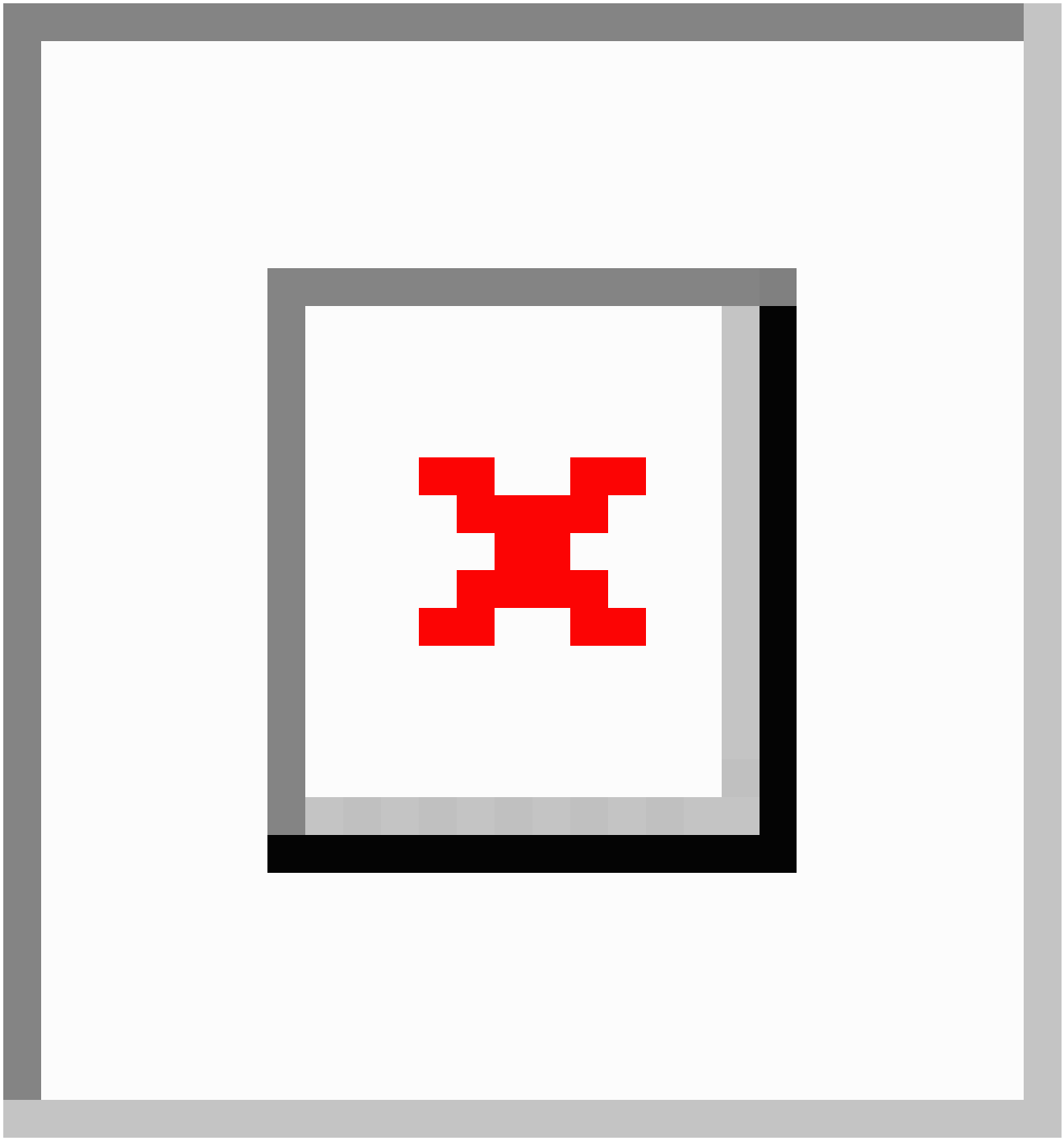
Pressure injuries comprise damage to skin and underlying tissue that usually occurs over a bony prominence but can be related to placement of medical devices [1]. The injury occurs because of intense or prolonged pressure that is combined with shear forces. Pressure injuries are a widespread and costly problem. A recent study found the prevalence of pressure injuries may be close to 30% for patients in intensive care units, which is 10% higher than previous estimates [2,3]. Patients with pressure injuries experience pain and the potential for infection and debilitation, which prolongs hospital stays and impacts recovery. Furthermore, increasing evidence supports the association between severity of pressure injuries and patient mortality [2]. In the United States, health care systems absorb on average US $10,000 per hospital-acquired pressure injury (HAPI), which contributes to a cost burden that will soon exceed US $30 billion [4,5].

Prevention of pressure injuries requires an accurate risk assessment and an interdisciplinary approach with routine repositioning, maintaining dry skin, and padding pressure points to reduce injury [6-8]. Currently, health care systems are striving to accurately measure and prevent HAPIs, since they can be common and negatively impact patient care [9]. Patient factors such as age, vasopressor support, mechanical ventilation, low

albumin, and renal failure can increase the risk for pressure injuries [10,11]. Multiple standardized risk assessment tools have been developed to systematically assess patient factors and assist clinicians in identifying at-risk patients [12,13]. Of these tools, the Braden scale has remained the standard of care across health systems for decades. The Braden scale incorporates components of sensory perception, activity, mobility, and nutrition, as well as skin moisture, friction, and shear force, to produce a score that indicates the risk of developing a pressure injury [14]. Although use of the Braden scale is widespread, its accuracy and reliability in diverse settings and patients is in question; thus, researchers have turned to more advanced risk prediction models that incorporate additional patient factors [12,13,15,16].

Recent literature reviews of advanced risk prediction models have highlighted excellent performance in predicting pressure injuries [17-21]. Zhou and colleagues [20] found that 74% of studies achieved an area under the receiver operating curve (AUC) between 0.68 and 0.99. Although these models were exceptionally accurate at predicting pressure injuries, no studies to our knowledge have implemented such models to reduce the number of pressure injuries. Numerous prediction models have been developed across clinical domains, but few have improved patient outcomes, leading researchers to identify a variety of required elements that may be necessary to implement prediction models in practice [22-24]. For instance, Randall Moorman [23] proposed properties, such as change of risk over time (eg, dynamic risk), for predictive analytics in neonatal intensive care units. Keim-Malpass and colleagues [24] found that potential users want prediction tools to be integrated with the electronic health record (EHR; eg, feasibility). We reviewed and agreed upon 5 elements that applied to HAPI prediction (ie, it should be feasible, broadly applicable, include dynamic risk and actionable criteria, and be rigorously validated) and then applied these elements to 22 recent models from 2020 to 2022 (Figure 1) [17,20,21]. We found no models fulfilled all the necessary elements to impact patient care. To help bridge the gap from model development to implementation, the objective of this study was, therefore, to develop and validate a model that fulfilled these elements and then compare its performance to usual care (ie, the Braden scale).

Figure 1. Comparison of current pressure injury prediction models according to elements of implementable models [25-45].



Methods

Study Population

We used retrospective data from the EHR at Vanderbilt University Medical Center between January 1, 2018, and July 1, 2022. All hospital admissions were included if the length of stay was longer than 24 hours and patient age was greater than 18 years on admission. HAPIs were identified using nurse flowsheet documentation. Nurses use flowsheets to document a variety of assessments, with our institution using a dedicated section for pressure injuries. The presence or absence of a pressure injury is assessed on admission and daily for each patient in the hospital. If a pressure injury is identified, the nurse documents whether it was present on admission and additional characteristics of the pressure injury, including the stage and location. We considered pressure injuries documented with a "no" in the column "present on admission" as HAPIs. For patients who had more than one HAPI, we used the first documented. The cohort included 197,911 hospitalizations, 129,100 patients, and 5458 HAPIs.

Feature Selection and Cohort Development

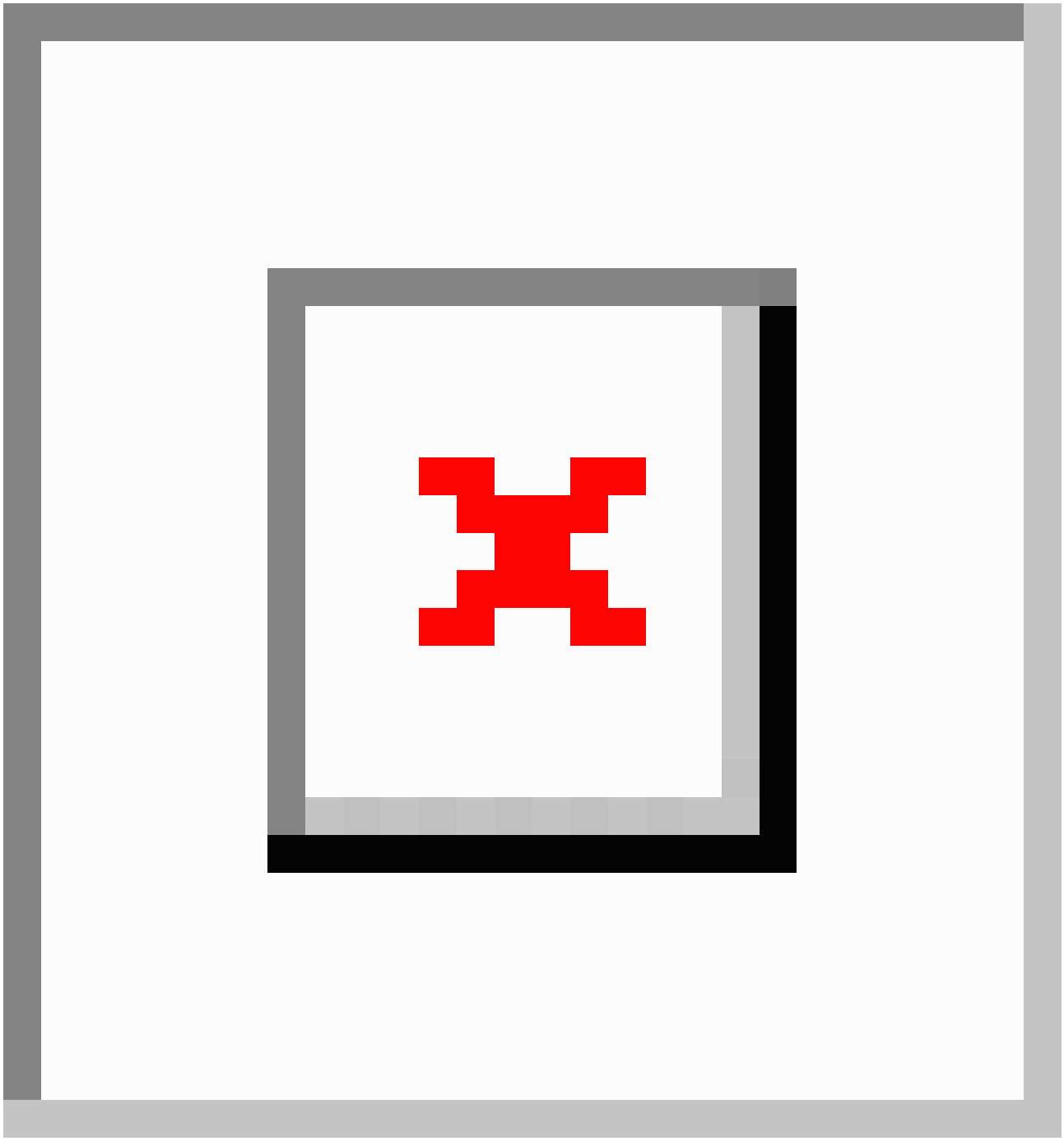
We first identified relevant features associated with pressure injuries from the literature. The list of relevant features was supplemented and pruned by clinical domain experts and informaticians at Vanderbilt University Medical Center. In total, 51 features were extracted as candidate features for predicting HAPIs. Importantly, features were only extracted if they were

available at the time of hospitalization and could be used to update the risk prediction during the encounter (ie, no claims data were used). Table 1 provides a summary of the extracted features. Missing values were imputed with the cohort median [46,47]. Multimedia Appendix 1 provides the full cohort characteristics, including missing values and a full list of measures. We split the full cohort temporally into model development and validation cohorts based on the number of events, with the development and validation cohorts including 80% and 20% of HAPIs, respectively. The development cohort included 161,816 hospitalizations and 4362 HAPIs from January 1, 2018, to August 26, 2021, and the validation cohort included 36,095 hospitalizations and 1096 HAPIs from August 27, 2021, to June 29, 2022 (Figure 2). Outcomes and features were identified and extracted in the same manner for the development and validation cohorts.

Table . Overview of extracted features.

Source	Feature
Patient demographics and social history	Age; gender; race; ethnicity; smoking status
Administration	Hospital admission through emergency department; intensive care unit admission; length of stay
Flowsheets	Hospital-acquired pressure injury (primary outcome); temperature; respiratory rate; heart rate; BMI; oxygen saturation; blood pressure; Braden scale (items and composite score); consciousness; gait transfer; Glasgow Coma Scale; malnutrition score; spinal cord injury; dialysis during hospitalization; tracheostomy; gastric tube; central line; chest tube; ostomy; drain; extracorporeal membrane oxygenation
Laboratory results	Hemoglobin; hemoglobin A_{1C}; hematocrit; mean corpuscular hemoglobin concentration; red cell distribution width; platelet count; chloride; blood urea nitrogen; creatinine; lactate; albumin; glucose

Figure 2. Model development and validation cohorts.



Model Development

We developed 3 models for comparison using logistic regression. The present model (Vanderbilt) used a broad set of candidate features (Table 1). The second model used the sum of the individual item measures from the Braden scale (ie, continuous Braden) [14]. Finally, since the Braden scale is typically operationalized using a single composite score (ie, less than 18=high risk; greater than or equal to 18=low risk), we included the dichotomous Braden for comparison as well. Logistic regression is the most frequently used model in clinical care [20,48]. The primary advantages of using logistic regression are that feature importance is easily interpretable and that the mathematical equation used to extract features and calculate a risk prediction is readily available in most commercial EHRs. Currently, the output from many machine learning models is not operationalizable for patient care in the EHR. To account for nonlinearity of the numeric features, we tested 3 knot-restricted cubic splines but found the discrimination failed to improve by using the nonlinear model [49]. Since the purpose was to develop a model that could be easily implemented in the EHR and compare it to standard care, we focused on use of logistic regression for the Vanderbilt and continuous Braden models.

We first included all 51 candidate features in the present (Vanderbilt) model to examine complexity versus accuracy as measured by cross-validation AUC. Again, included features were derived from the literature and refined by clinical domain

experts and informaticians. We tested for multicollinearity by examining the proportion of variance in each candidate feature that could be explained by other candidate features and removed hemoglobin. Included features had to be structured and readily available for automated processing in the EHR without additional input by the user. Using the conservative 15:1 rule, we were able to include 290.8 degrees of freedom in the model. To ensure the model was broadly applicable across settings and patients, we used a least absolute shrinkage and selection operator (LASSO) approach to identify important candidate features. Candidate features were standardized (scaled and centered) prior to running the LASSO regression. LASSO introduces a penalty term to the standard regression model, which forces some of the regression coefficients to shrink toward zero, effectively performing feature selection [50]. Variables with nonzero coefficients were included in the final model. The model was designed to calculate a risk prediction on admission and daily while the patient was in the hospital. Missing numeric measures were to be imputed with the cohort median until measures became available.

Model Evaluation

The final model was assessed in an external cohort that was temporally separated from the model development cohort. We evaluated the model using traditional and novel performance measures, which included the AUC, Brier score, slope, intercept, integrated calibration index, and calibration curve. AUC is a performance measure for the discrimination of HAPI versus no HAPI. It combines the true and false positive rates, with an AUC of 0.5 indicating no meaningful discrimination. The Brier score accounts for the predicted HAPI outcome as well as the estimate and is calculated by the squared difference between the prediction (0 to 1) and outcome (0=no HAPI and 1=HAPI) [51]. For example, if a patient had a 90% probability of developing a HAPI and did develop a HAPI during that encounter, the Brier score would be 0.01. A Brier score of 0 indicates perfect accuracy and a score of 1 indicates perfect inaccuracy. The integrated calibration index is a numeric summary of model calibration across the predicted probabilities [52]. It is the weighted average of the absolute difference between the observed and predicted probabilities; therefore, a lower integrated calibration index indicates better calibration. A slope equal to 1 indicates agreement between the observed response and the predicted probability, while a slope greater than 1 indicates potential underfitting, and a slope lower than 1 indicates potential overfitting [52]. Similarly, an intercept of zero is ideal. As with prior models, no adjustments were made for multiple comparisons [47,53,54]. We used the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) reporting guidelines (Checklist 1) and performed all analyses in R (version 4.2.3; R Foundation for Statistical Computing) with relevant extension packages [55].

Ethical Considerations

This study was approved by the Vanderbilt University Medical Center Institutional Review Board (220644), and data were deidentified.

Results

Cohort Characteristics

The full cohort of patient encounters was split temporally, based on the number of HAPIs, into model development and validation cohorts. The characteristics for each cohort are provided in Table 2. Among the model development cohort, those who developed HAPIs were older and male. Table 3 provides the model development cohort characteristics divided by whether a HAPI occurred.

Table . Characteristics for model development and validation cohorts. Measures were first taken during the hospital stay unless specified otherwise. Race and ethnicity were not included as candidate features.

		Development cohort (n=161,816 encounters)	Validation cohort (n=36,095 encounters)
Age (years), median (IQR)		56 (37-69)	56 (37-69)
Female, n (%)		84,727 (52.4)	17,060 (47.3)
Race, n (%)			
	White	125,322 (77.4)	27,649 (76.6)
	African American	26,299 (16.3)	5659 (15.7)
	Asian	2325 (1.4)	480 (1.3)
	American Indian or Alaska Native	289 (0.2)	66 (0.2)
	Pacific Islander	104 (0.1)	25 (0.1)
	Multiple	1185 (0.7)	231 (0.6)
Hispanic, n (%)		7406 (4.6)	2074 (5.7)
Physiological and clinical features			
	Temperature (°C), median (IQR)	36.7 (36.5-36.9)	36.7 (36.5-36.9)
	Respiratory rate (breaths per minute), median (IQR)	18 (16-19)	18 (16-20)
	Heart rate (beats per minute), median (IQR)	87 (74-100)	87 (75-101)
	BMI (kg/m^2), median (IQR)	28.2 (24.1-33.4)	28.3 (24.2-33.5)
	Oxygen saturation (%), median (IQR)	98 (96-99)	98 (96-99)
	Systolic blood pressure (mm Hg), median (IQR)	131 (117-147)	130 (117-146)
	Diastolic blood pressure (mm Hg), median (IQR)	77 (68-88)	77 (67-87)
	Emergency department admissions, n (%)	91,363 (56.5)	20,972 (58.1)
	Intensive care admissions, n (%)	34,190 (21.1)	7831 (21.7)
	Length of stay (days), median (IQR)	4 (2-6)	4 (2-7)
	Smokers, n (%)	56,750 (35.1)	12,561 (34.8)
	Edema, n (%)	86,582 (53.5)	19,846 (55)
	Spinal cord injury, n (%)	5908 (3.7)	1428 (4)
	Dialysis, n (%)	92 (0.1)	74 (0.2)
	Tracheostomy, n (%)	2122 (1.3)	520 (1.4)
	Gastric tube, n (%)	35 (0)	5 (0)
	Central line, n (%)	20,648 (12.8)	4803 (13.3)
	Chest tube, n (%)	5186 (3.2)	1278 (3.5)
	Ostomy, n (%)	2059 (1.3)	459 (1.3)
	Drain, n (%)	17,800 (11)	4005 (11.1)
	ECMO[a], n (%)	414 (0.3)	71 (0.2)
Laboratory results, median (IQR)			
	Hemoglobin A_{1C} (%)	6.1 (5.5-7.5)	6.1 (5.6-7.5)
	Hemoglobin (g/dL)	12.0 (10.3-13.6)	11.9 (10.2-13.5)
	Hematocrit (%)	36.0 (32.0-41.0)	36.0 (32.0-40.0)
	MCHC[b] (g/dL)	33.0 (32.0-34.0)	32.9 (31.9-33.9)
	Red cell distribution width (%)	13.9 (13.0-15.5)	14.0 (13.0-15.6)
	Platelet count ($\times 10^9$/L)	228 (174-291)	234 (179-298)
	Chloride (mEq/L)	105 (101-108)	104 (101-107)

		Development cohort (n=161,816 encounters)	Validation cohort (n=36,095 encounters)
	Lactate (mmol/L)	1.1 (0.8-1.9)	1.2 (0.8-2.0)
	Albumin (g/dL)	3.6 (3.1-4.0)	3.5 (3.0-3.9)
	Urine blood urea nitrogen	412 (260-603)	415 (275-609)
	Creatinine (mg/dL)	0.9 (0.9-1.3)	0.9 (0.8-1.3)
	Glucose (mmol/L)	114 (96-146)	114 (96-145)
Nursing assessment features			
	Braden scale score, median (IQR)	20 (18-22)	20 (17-21)
	Level of consciousness=2, n (%)	21,357 (13.2)	5043 (14)
	Gait transfer=20, n (%)	10,673 (6.6)	2190 (6.1)
	Glasgow Coma Scale=3, n (%)	4872 (3)	961 (2.7)
	Malnutrition score=5, n (%)	1241 (0.8)	345 (1)
Outcomes, n (%)			
	Any pressure injury	9259 (5.7)	2143 (5.9)
	Hospital-acquired pressure injury	4362 (2.7)	1096 (3)

[a]ECMO: extracorporeal membrane oxygenation.

[b]MCHC: mean corpuscular hemoglobin concentration.

Table . Model development cohort characteristics with and without hospital acquired pressure injury. Measures were the first taken during the hospital stay unless specified otherwise. Race and ethnicity were not included as candidate features.

		No hospital-acquired pressure injury (n=157,454 encounters)	Hospital-acquired pressure injury (n=4362 encounters)
Age (years), median (IQR)		56 (37-68)	64 (52-74)
Female, n (%)		82,999 (52.7)	1728 (39.6)
Race, n (%)			
	White	121,786 (77.3)	3536 (81.1)
	African American	25,654 (16.3)	645 (14.8)
	Asian	2290 (1.5)	35 (0.8)
	American Indian or Alaska Native	285 (0.2)	4 (0.1)
	Pacific Islander	102 (0.1)	2 (0)
	Multiple	1154 (0.7)	31 (0.7)
Hispanic, n (%)		7306 (4.6)	100 (2.3)
Physiologic and clinical features			
	Temperature (°C), median (IQR)	36.7 (36.5-36.9)	36.7 (36.4-37.0)
	Respiratory rate (breaths per minute), median (IQR)	18.0 (16.0-19.0)	18.0 (16.0-22.0)
	Heart rate (beats per minute), median (IQR)	87.0 (74.0-100.0)	91.0 (77.0-106.0)
	BMI (kg/m^2), median (IQR)	28.2 (24.1-33.5)	26.8 (22.6-32.2)
	Oxygen saturation (%), median (IQR)	98.0 (96.0-99.0)	97.0 (95.0-99.0)
	Systolic blood pressure (mm Hg), median (IQR)	131.0 (117.0-147.0)	124.0 (107.0-142.0)
	Diastolic blood pressure (mm Hg), median (IQR)	78.0 (68.0-88.0)	72.0 (61.0-84.0)
	Emergency department admissions, n (%)	88,552 (56.2)	2811 (64.4)
	Intensive care admissions, n (%)	31,795 (20.2)	2395 (54.9)
	Length of stay (days), median (IQR)	3 (2-6)	15 (8-26)
	Smokers, n (%)	55,278 (35.1)	1472 (33.7)
	Edema, n (%)	82,640 (52.5)	3942 (90.4)
	Spinal cord injury, n (%)	5398 (3.4)	510 (11.7)
	Dialysis, n (%)	81 (0.1)	11 (0.3)
	Tracheostomy, n (%)	1491 (0.9)	631 (14.5)
	Gastric tube, n (%)	24 (0)	11 (0.3)
	Central line, n (%)	18,350 (11.7)	2298 (52.7)
	Chest tube, n (%)	4598 (2.9)	588 (13.5)
	Ostomy, n (%)	1881 (1.2)	178 (4.1)
	Drain, n (%)	16,888 (10.7)	912 (20.9)
	ECMO[a], n (%)	242 (0.2)	172 (3.9)
Laboratory results, median (IQR)			
	Hemoglobin A_{1C} (%)	6.1 (5.5-7.5)	6.0 (5.4-7.1)
	Hemoglobin (g/dL)	12.0 (10.3-13.6)	12.0 (10.3-13.6)
	Hematocrit (%)	37.0 (32.0-41.0)	34.0 (29.0-40.0)

		No hospital-acquired pressure injury (n=157,454 encounters)	Hospital-acquired pressure injury (n=4362 encounters)
	MCHC[b] (g/dL)	33.0 (32.0-34.0)	32.6 (31.5-33.7)
	Red cell distribution width (%)	13.9 (13.0-15.5)	14.9 (13.5-16.8)
	Platelet count ($\times 10^9$/L)	228.0 (175.0-291.0)	215.0 (151.0-298.0)
	Chloride (mEq/L)	105.0 (101.0-108.0)	104.0 (99.0-108.0)
	Lactate (mmol/L)	1.1 (0.8-1.3)	1.4 (0.9-2.5)
	Albumin (g/dL)	3.6 (3.1-4.0)	3.1 (2.6-3.5)
	Urine blood urea nitrogen	412.0 (263.0-603.0)	410.0 (244.0-605.5)
	Creatinine (mg/dL)	0.9 (0.8-1.3)	1.2 (0.8-1.9)
	Glucose (mmol/L)	114.0 (96.0-145.0)	125.0 (101.0-168.0)
Nursing assessment features			
	Braden scale score, median (IQR)	20.0 (18.0-22.0)	15.0 (13.0-18.0)
	Level of consciousness=2, n (%)	20,712 (13.2)	645 (14.8)
	Gait transfer=20, n (%)	10,055 (6.4)	618 (14.2)
	Glasgow Coma Scale=3, n (%)	4406 (2.8)	466 (10.7)
	Malnutrition score=5, n (%)	1166 (0.7)	75 (1.7)

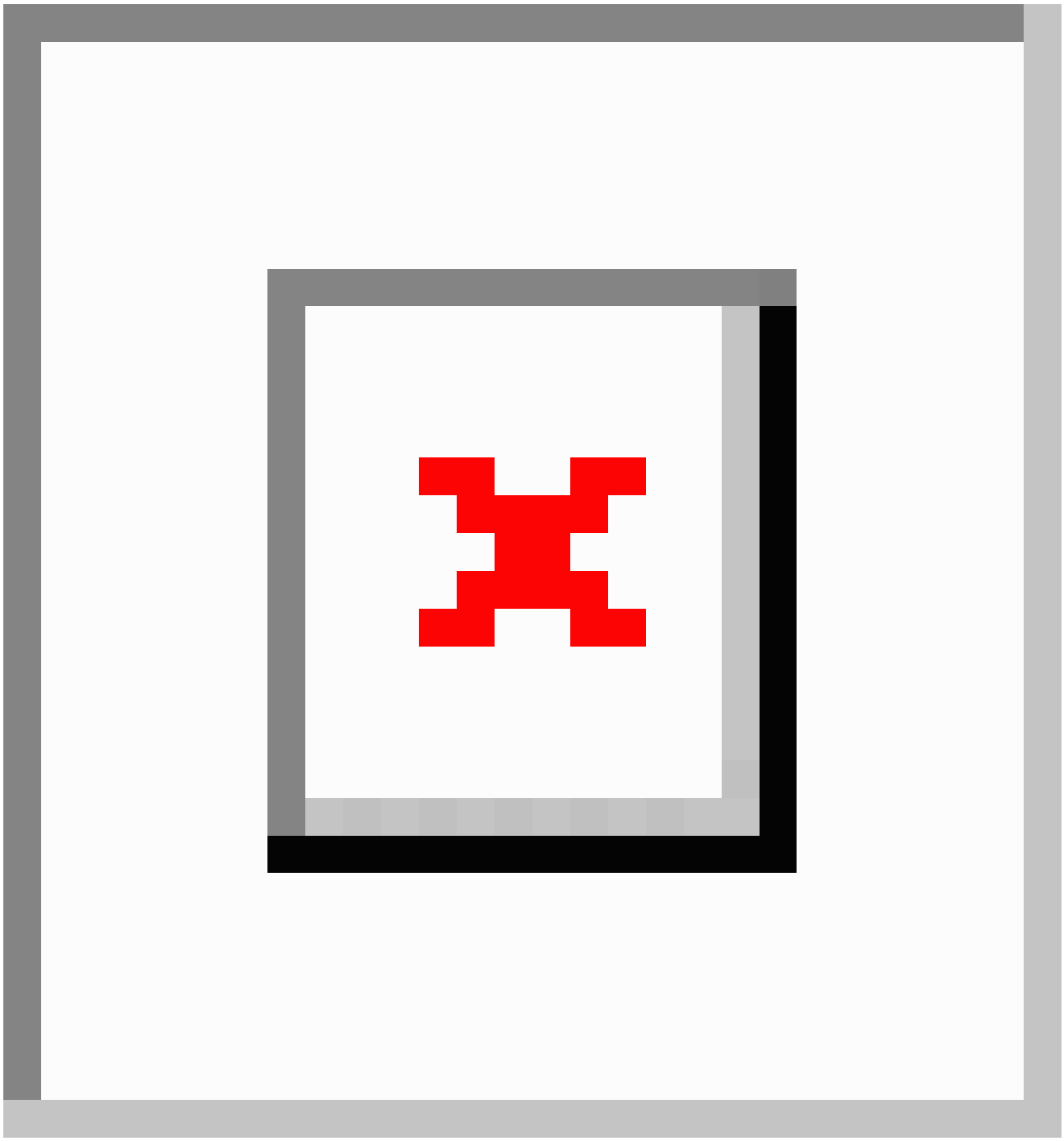
[a]ECMO: extracorporeal membrane oxygenation.

[b]MCHC: mean corpuscular hemoglobin concentration.

Model Description

We determined 22 features were necessary to achieve a parsimonious yet highly accurate model. Again, features were selected using a LASSO approach. We fit the final model with 4362 HAPI encounters and 291 degrees of freedom, which indicated the model was unlikely to overfit the data. Of the 40 features that exhibited association with developing a HAPI, the top 5 features included tracheostomy (odds ratio [OR] 4.5, 95% CI 4.0-5.1), peripheral edema (OR 2.9, 95% CI 2.6-3.2), central line (OR 2.1, 95% CI 1.9-2.3), first albumin measure (OR 0.6, 95% CI 0.6-0.6), and age (OR 1.2, 95% CI 1.2-1.2) (Figure 3). Although the directionality for each feature may vary, the relative importance in Figure 3 was ranked on a single scale. Additional significant features included whether the patient was on sympathomimetic medications, had a spinal cord injury or chest tube, and individual Braden score component measures. The final Vanderbilt model with 22 features provided excellent discriminatory ability with an AUC of 0.897 (95% CI 0.893-0.901). Multimedia Appendix 2 depicts the probability density plot for the development and validation cohorts.

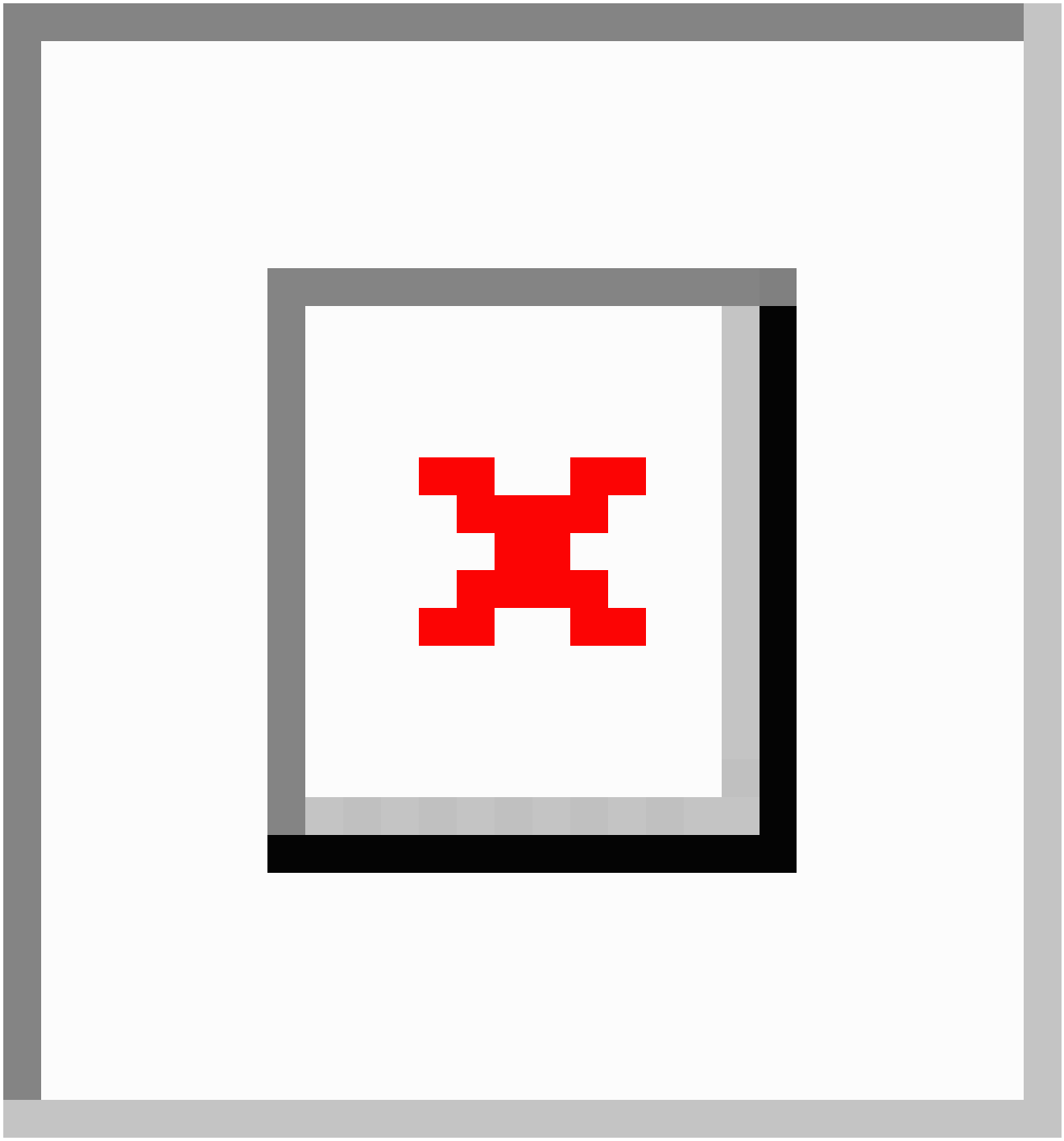
Figure 3. Relative importance of features used in the final Vanderbilt model. Gray subfeatures represent item comparisons used to generate features. *P* values for variable significance were derived using the Wald χ^2 test. BUN: blood urea nitrogen; ECMO: extracorporeal membrane oxygenation; ICU: intensive care unit; RDW: red cell distribution width.



Comparison With the Braden Scale

Using the model development cohort, the Vanderbilt model achieved an AUC of 0.897 (95% CI 0.893-0.901), compared to 0.798 (95% CI 0.791-0.803) and 0.733 (95% CI 0.725-0.740) for the continuous and dichotomous Braden, respectively (Figure 4).

Figure 4. Area under the receiver operating characteristic curve comparing the Vanderbilt (gold), continuous Braden (blue), and dichotomous Braden (gray) models.



Model Validation

The validation cohort consisted of 34,999 hospitalizations without a HAPI and 1096 hospitalizations with at least one HAPI. Model development and validation cohorts were compared to confirm that each had similar characteristics. Overall, characteristics were similar between the 2 cohorts (Table 3). We applied the same model from the development cohort to the validation cohort without adjusting coefficients, which provided a concordance statistic of 0.893 (95% CI 0.885-0.899; Table 4). Model calibration was consistent between the development and validation cohorts. The calibration curve indicated the model most accurately predicted risk for patients in the range of 0%-25% predicted risk (Figure 5); above this, the model could overpredict a HAPI. Since the model was intended to bring nurse attention and interventions to patients who would otherwise be overlooked, we believe the miscalibration at higher percentages was less clinically relevant. There was no evidence of collinearity. We are confident that this model performs well for most patients across the intensive care and general hospital settings, as 98.2% of the cohort had a predicted risk of less than 25%.

Since the model was designed to be used broadly in the general adult hospital, we performed a post hoc analysis among subpopulations for age (older than 65 years), gender, race, ethnicity, intensive care unit admission, and Braden score (greater than 18). The subpopulation analysis revealed only

slight changes in discrimination performance (Multimedia Appendix 3).

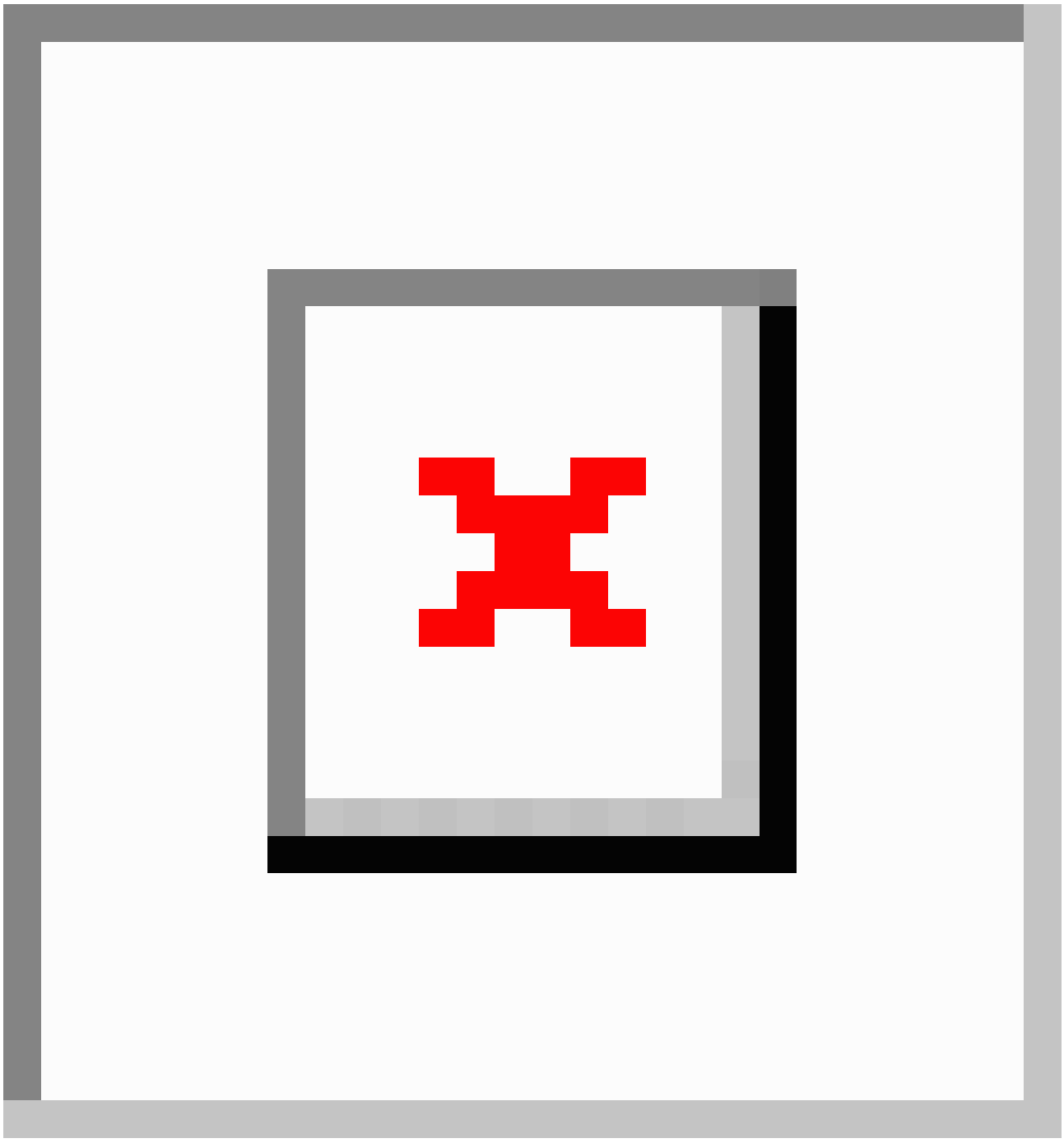
To operationalize the Vanderbilt model in the EHR (Epic), we generated the equation below. The output from the equation is a numeric probability from 0 to 1. Z is the sum of –4.1812002 and the product of the coefficient and measured value (eg, first albumin) for each feature. In Multimedia Appendix 4, we provide the coefficients for the equation. The model has been deployed as a population management tool to generate risk prediction data at Vanderbilt University Medical Center, but the output is only available for the research team until a trial period has been completed and governance has approved it for patient care. Within a report for multiple patients, output from the model is available as a column among other relevant factors to prioritize pressure injury interventions. As part of the implementation plan, we have created an application for potential users to test the model [56].

$$\text{Probability of hospital-acquired pressure injury} = \frac{1}{(1+\exp(-Z))}$$

Table . Prediction model performance for hospital-acquired pressure injury.

Model	Area under the curve (95% CI)	Brier score	Integrated calibration index	Intercept	Slope
Vanderbilt (logistic regression)	0.893 (0.885-0.899)	0.026	0.006	−0.041	0.977
Continuous Braden (logistic regression)	0.799 (0.789-0.811)	0.028	0.006	0.178	1.034
Dichotomous Braden (score<18)	0.733 (0.725-0.740)	0.025	Too few levels to compute	0.0	1.0

Figure 5. Calibration curves for model development (left) and validation (right). Logistic calibration (solid line) represents parameter-based calibration (logistic regression model fit between predicted and observed values). Nonparametric calibration (dotted line) represents locally estimated scatterplot smoothing trend between predicted and observed values.



Discussion

Principal Findings

We developed and validated a risk prediction model for HAPIs that can be used in the general adult population. The model achieved excellent discrimination and adequate calibration (Table 4). Although several recent models have achieved similar performance, our model may have the greatest likelihood of reducing HAPIs because it was built with the foresight of overcoming known barriers to implementation of risk-prediction clinical decision support (Figure 1). According to the scoring criteria in Figure 1, the present model would have achieved 8 of a possible 10, compared to the current highest score of 6. It lost points for being limited to adults from a single institution (broadly applicable) and partially specified intervention (actionable criteria). Limiting development of the model to a single institution could limit the generalizability due to documentation patterns and data availability. Although we specified how to deploy the model in the EHR, the intervention components and implementation strategies were underspecified for implementation and evaluation. The next step is to test the effectiveness of the model in a pragmatic randomized clinical trial in which the intervention will be fully specified [57].

Although our model achieved similar performance and used the same regression approach as the top 3 models in Figure 1 (Ladios-Martin et al [25], Levy et al [27], and Song et al [26]), many of the most important features among the models varied. Among the most important features in the Ladios-Martin et al [25] model (eg, medical service, days of antidiabetic therapy, ability to eat, number of red blood cell units transfused, and hemoglobin range), only medical service was similar to our model. Relatedly, 2 important features in the Levy et al [27] model overlapped (friction and mobility). However, several important features from the Song et al [26] model (albumin, gait/transferring, activity, blood urea nitrogen, chloride, and spinal cord injury) overlapped with our model. We anticipate the similarity in features between our model and the Song et al [26] model was due to use of the same EHR and the models being developed at academic medical centers in the United States.

Limited implementation of risk prediction models in the EHR presents a critical challenge in health care today; the barrier is now less about the performance of risk prediction models and more the sociotechnical obstacles to uptake in patient care [58-60]. Despite the growing availability and sophistication of these models, their integration into routine clinical practice remains inadequate. Of the 22 models identified, we were unable to find one that decreased HAPIs. Even when prespecified elements for an implementable model are fulfilled, concerted efforts are needed from various stakeholders. Collaboration between health care organizations, technology developers, and regulatory bodies is essential to establish standards and guidelines for incorporating risk prediction models into EHR systems [61]. Enhancing data infrastructure, promoting data standardization, and developing robust privacy and security frameworks are crucial steps toward facilitating the implementation of these models [62]. Additionally, targeted education and training initiatives can help build trust and confidence among health care providers, encouraging their acceptance and use of risk prediction models in clinical practice, along with actionable steps to take for patients at highest risk [63,64]. Furthermore, there are significant socio-organizational barriers that impede the implementation of risk prediction models in EHRs. Resistance to change, lack of awareness or understanding among health care providers, and concerns regarding liability and accountability are common challenges faced by health care institutions. Clinicians may be skeptical of relying on risk prediction models, fearing that their judgment and decision-making autonomy may be compromised. The integration of risk prediction models also requires extensive training and education for health care providers, which may be resource-intensive and time-consuming [65,66]. Only when these barriers are addressed in a pragmatic manner can risk-prediction clinical decision support models improve patient outcomes.

Pragmatic trials are crucial in testing the real-world effectiveness and utility of interventions in health care settings [57,67,68]. These trials provide valuable insights into how interventions perform when integrated into routine clinical practice, considering factors such as patient outcomes, workflow integration, and usability. Institutions are beginning to develop the infrastructure and stakeholder engagement to support pragmatic trials. At our institution, Semler and colleagues [69] tested the effectiveness of balanced crystalloids and saline for fluids in critically ill adults. This pragmatic trial was cluster-randomized with 5 intensive care units. The authors found that use of balanced crystalloids resulted in a lower rate of death. A key aspect that makes pragmatic trials feasible is the use of existing infrastructure and real-world practice, which typically includes an inclusive patient population, minimal staff training, flexible protocols, minimally disruptive interventions, and outcomes captured as part of care. For pressure injuries specifically, the intervention infrastructure and guidance already exist as part of routine care; however, risk prediction will help identify and prioritize the most at-risk patients for targeted intervention. Preliminarily, we envision a clinician will use a list of patients ranked highest to lowest risk for HAPI.

Strengths and Limitations

Pressure injury prediction models have shown promise in identifying individuals at risk of developing pressure injuries. However, there are several limitations with these models, including ours, that should be considered. First, documentation of pressure injuries varies by institution and can lead to misclassification. We found that documentation of some pressure injuries carried over from previous encounters. On further testing, we found that missing measures (eg, albumin) can lead to inaccurate prediction. Thus, we chose to use a replicable imputation method with the median. Although our prediction model was developed and validated using incident HAPIs, documentation errors should be carefully considered. To increase the generalizability of our model, we chose not to include text from notes, despite evidence that use of clinical notes may have predictive power. Although we had a relatively large sample size that was sufficient to include all important features, the patient cohort was from a single institution and may not generalize to institutions in different geographical areas or using different EHRs. Finally, we chose to use an interpretable model that could be operationalized in current EHRs; however, other models may provide slightly higher performance. We anticipate certain EHR vendors will continue to develop capabilities for implementing complex machine learning models for more complicated prediction tasks. In anticipation of this, we performed a preliminary analysis of random forest, generalized additive model, and XGBoost. Of these models, we found that XGBoost had higher discrimination than ours in the model development cohort (AUC 0.960, 95% CI 0.957-0.962 vs AUC 0.893, 95% CI 0.885-0.899). In the model validation cohort, however, performance was not superior to logistic regression (AUC 0.869, 95% CI 0.861-0.877 vs AUC 0.893, 95% CI 0.885-0.899). Future work is needed to fully optimize the machine learning models and explore the tradeoff between interpretability and performance.

Conclusion

Despite numerous models developed to predict pressure injuries, studies demonstrating improved patient outcomes are missing. This is because implementing risk prediction models for routine patient care is complex and requires model developers, clinicians, and researchers to address challenges early in the

process. Therefore, we developed and validated an accurate prediction model for HAPIs that fulfilled necessary elements for implementation. The next step is to overcome socio-organizational barriers to rigorously evaluate the model through a pragmatic randomized clinical trial that includes targeted intervention for patients at highest risk. Our approach to developing an implementable risk prediction model, with feasible plans to evaluate its effectiveness, is generalizable to risk prediction and may be necessary to unlock the potential of this technology and improve decision-making.

Acknowledgments

We would like to thank Donald Sengstack for helping with the data extraction and Lance Mailloux for guidance on pressure injury quality improvement. This study was funded by the National Institutes of Health (R01 AG062499) and the Advanced Vanderbilt Artificial Intelligence Laboratory.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full cohort characteristics.
[DOCX File, 33 KB - medinform_v12i1e51842_app1.docx]

Multimedia Appendix 2
Density and probability plots for model development (A) and validation (B).
[PNG File, 51 KB - medinform_v12i1e51842_app2.png]

Multimedia Appendix 3
Subpopulation analysis of adult general hospital patients.
[DOCX File, 18 KB - medinform_v12i1e51842_app3.docx]

Multimedia Appendix 4
Features and coefficients of model and equation.
[DOCX File, 16 KB - medinform_v12i1e51842_app4.docx]

Checklist 1
Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist.
[PDF File, 115 KB - medinform_v12i1e51842_app5.pdf]

References

1. Edsberg LE, Black JM, Goldberg M, McNichol L, Moore L, Sieggreen M. Revised National Pressure Ulcer Advisory Panel pressure injury staging system. J Wound Ostomy Continence Nurs 2016;43(6):585-597. [doi: 10.1097/WON.0000000000000281] [Medline: 27749790]
2. Labeau SO, Afonso E, Benbenishty J, et al. Prevalence, associated factors and outcomes of pressure injuries in adult intensive care unit patients: the DecubICUs study. Intensive Care Med 2021 Feb;47(2):160-169. [doi: 10.1007/s00134-020-06234-9] [Medline: 33034686]
3. Mervis JS, Phillips TJ. Pressure ulcers: pathophysiology, epidemiology, risk factors, and presentation. J Am Acad Dermatol 2019 Oct;81(4):881-890. [doi: 10.1016/j.jaad.2018.12.069] [Medline: 30664905]
4. Padula WV, Mishra MK, Makic MBF, Sullivan PW. Improving the quality of pressure ulcer care with prevention: a cost-effectiveness analysis. . 2011 Apr(4) p. 385-392. [doi: 10.1097/MLR.0b013e31820292b3] [Medline: 21368685]
5. Padula WV, Delarmente BA. The national cost of hospital-acquired pressure injuries in the United States. Int Wound J 2019 Jun;16(3):634-640. [doi: 10.1111/iwj.13071] [Medline: 30693644]
6. Wound, Ostomy and Continence Nurses Society-Wound Guidelines Task Force. WOCN 2016 Guideline for Prevention and Management of Pressure Injuries (Ulcers): an executive summary. J Wound Ostomy Continence Nurs 2017;44(3):241-246. [doi: 10.1097/WON.0000000000000321] [Medline: 28472816]
7. Reddy M, Gill SS, Kalkar SR, Wu W, Anderson PJ, Rochon PA. Treatment of pressure ulcers: a systematic review. JAMA 2008 Dec 10;300(22):2647-2662. [doi: 10.1001/jama.2008.778] [Medline: 19066385]
8. Reddy M, Gill SS, Rochon PA. Preventing pressure ulcers: a systematic review. JAMA 2006 Aug 23;296(8):974-984. [doi: 10.1001/jama.296.8.974] [Medline: 16926357]

9. Kavanagh KT, Dykes PC. Hospital pressure injury metrics, an unfulfilled need of paramount importance. J Patient Saf 2021 Apr 1;17(3):189-191. [doi: 10.1097/PTS.0000000000000694] [Medline: 32805091]
10. Alderden J, Rondinelli J, Pepper G, Cummins M, Whitney J. Risk factors for pressure injuries among critical care patients: a systematic review. Int J Nurs Stud 2017 Jun;71:97-114. [doi: 10.1016/j.ijnurstu.2017.03.012] [Medline: 28384533]
11. Serrano ML, Méndez MIG, Cebollero FMC, Rodríguez JSL. Risk factors for pressure ulcer development in intensive care units: a systematic review. Med Intensiva 2017 Aug;41(6):339-346. [doi: 10.1016/j.medine.2017.04.006]
12. Liao Y, Gao G, Mo L. Predictive accuracy of the Braden Q scale in risk assessment for paediatric pressure ulcer: a meta-analysis. Int J Nurs Sci 2018 Oct 10;5(4):419-426. [doi: 10.1016/j.ijnss.2018.08.003] [Medline: 31406858]
13. Wei M, Wu L, Chen Y, Fu Q, Chen W, Yang D. Predictive validity of the Braden scale for pressure ulcer risk in critical care: a meta-analysis. Nurs Crit Care 2020 May;25(3):165-170. [doi: 10.1111/nicc.12500] [Medline: 31985893]
14. Papanikolaou P, Lyne P, Anthony D. Risk assessment scales for pressure ulcers: a methodological review. Int J Nurs Stud 2007 Feb;44(2):285-296. [doi: 10.1016/j.ijnurstu.2006.01.015] [Medline: 17141782]
15. Huang C, Ma Y, Wang C, et al. Predictive validity of the Braden scale for pressure injury risk assessment in adults: a systematic review and meta-analysis. Nurs Open 2021 Sep;8(5):2194-2207. [doi: 10.1002/nop2.792] [Medline: 33630407]
16. Hyun S, Vermillion B, Newton C, et al. Predictive validity of the Braden scale for patients in intensive care units. Am J Crit Care 2013 Nov;22(6):514-520. [doi: 10.4037/ajcc2013991] [Medline: 24186823]
17. Dweekat OY, Lam SS, McGrath L. Machine learning techniques, applications, and potential future opportunities in pressure injuries (bedsores) management: a systematic review. Int J Environ Res Public Health 2023 Jan 1;20(1):796. [doi: 10.3390/ijerph20010796] [Medline: 36613118]
18. Jiang M, Ma Y, Guo S, et al. Using machine learning technologies in pressure injury management: systematic review. JMIR Med Inform 2021 Mar 10;9(3):e25704. [doi: 10.2196/25704] [Medline: 33688846]
19. Ribeiro F, Fidalgo F, Silva A, Metrôlho J, Santos O, Dionisio R. Literature review of machine-learning algorithms for pressure ulcer prevention: challenges and opportunities. Informatics 2021 Dec 1;8(4):76. [doi: 10.3390/informatics8040076]
20. Zhou Y, Yang X, Ma S, Yuan Y, Yan M. A systematic review of predictive models for hospital-acquired pressure injury using machine learning. Nurs Open 2023 Mar;10(3):1234-1246. [doi: 10.1002/nop2.1429] [Medline: 36310417]
21. Qu C, Luo W, Zeng Z, et al. The predictive effect of different machine learning algorithms for pressure injuries in hospitalized patients: a network meta-analyses. Heliyon 2022 Nov;8(11):e11361. [doi: 10.1016/j.heliyon.2022.e11361] [Medline: 36387440]
22. Kitzmiller RR, Vaughan A, Skeeles-Worley A, et al. Diffusing an innovation: clinician perceptions of continuous predictive analytics monitoring in intensive care. Appl Clin Inform 2019 Mar;10(2):295-306. [doi: 10.1055/s-0039-1688478] [Medline: 31042807]
23. Randall Moorman J. The principles of whole-hospital predictive analytics monitoring for clinical medicine originated in the neonatal ICU. NPJ Digit Med 2022 Mar 31;5(1):41. [doi: 10.1038/s41746-022-00584-y] [Medline: 35361861]
24. Keim-Malpass J, Kitzmiller RR, Skeeles-Worley A, et al. Advancing continuous predictive analytics monitoring: moving from implementation to clinical action in a learning health system. Crit Care Nurs Clin North Am 2018 Jun;30(2):273-287. [doi: 10.1016/j.cnc.2018.02.009] [Medline: 29724445]
25. Ladios-Martin M, Fernández-de-Maya J, Ballesta-López FJ, Belso-Garzas A, Mas-Asencio M, Cabañero-Martínez MJ. Predictive modeling of pressure injury risk in patients admitted to an intensive care unit. Am J Crit Care 2020 Jul 1;29(4):e70-e80. [doi: 10.4037/ajcc2020237] [Medline: 32607572]
26. Song W, Kang MJ, Zhang L, 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. [doi: 10.1093/jamia/ocaa336] [Medline: 33517452]
27. Levy JJ, Lima JF, Miller MW, Freed GL, O'Malley AJ, Emeny RT. Machine learning approaches for hospital acquired pressure injuries: a retrospective study of electronic medical records. Front Med Technol 2022 Jun;4:926667. [doi: 10.3389/fmedt.2022.926667] [Medline: 35782577]
28. Ossai CI, O'Connor L, Wickramasinghe N. Real-time inpatients risk profiling in acute care: a comparative study of falls and pressure injuries vulnerabilities. In: Pucihar A, Kljajic Borstnar M, Bons R, editors. 33rd BLED eConference: Enabling Technology for a Sustainable Society: University of Maribor Press; 2021:35-50. [doi: 10.18690/978-961-286-362-3.3]
29. Cai JY, Zha ML, Song YP, Chen HL. Predicting the development of surgery-related pressure injury using a machine learning algorithm model. J Nurs Res 2020 Dec 21;29(1):e135. [doi: 10.1097/JNR.0000000000000411] [Medline: 33351552]
30. Xu J, Chen D, Deng X, et al. Development and validation of a machine learning algorithm–based risk prediction model of pressure injury in the intensive care unit. Int Wound J 2022 Nov;19(7):1637-1649. [doi: 10.1111/iwj.13764] [Medline: 35077000]
31. Šín P, Hokynková A, Marie N, Andrea P, Krč R, Podroužek J. Machine learning-based pressure ulcer prediction in modular critical care data. Diagnostics (Basel) 2022 Mar 30;12(4):850. [doi: 10.3390/diagnostics12040850] [Medline: 35453898]
32. Do Q, Lipatov K, Ramar K, Rasmusson J, Pickering BW, Herasevich V. Pressure injury prediction model using advanced analytics for at-risk hospitalized patients. J Patient Saf 2022 Oct 1;18(7):e1083-e1089. [doi: 10.1097/PTS.0000000000001013] [Medline: 35588068]

33. Walther F, Heinrich L, Schmitt J, Eberlein-Gonska M, Roessler M. Prediction of inpatient pressure ulcers based on routine healthcare data using machine learning methodology. Sci Rep 2022 Mar 23;12(1):5044. [doi: 10.1038/s41598-022-09050-x] [Medline: 35322109]
34. Anderson C, Bekele Z, Qiu Y, Tschannen D, Dinov ID. Modeling and prediction of pressure injury in hospitalized patients using artificial intelligence. BMC Med Inform Decis Mak 2021 Aug 30;21(1):253. [doi: 10.1186/s12911-021-01608-5] [Medline: 34461876]
35. Cheng FM, Jin YJ, Chien CW, Chuang YC, Tung TH. The application of Braden scale and rough set theory for pressure injury risk in elderly male population. J Mens Health 2021 Sep;17(4):156-165. [doi: 10.31083/jomh.2021.022]
36. Song J, Gao Y, Yin P, et al. The random forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms. Risk Manag Healthc Policy 2021 Mar;14:1175-1187. [doi: 10.2147/RMHP.S297838] [Medline: 33776495]
37. Nakagami G, Yokota S, Kitamura A, et al. Supervised machine learning-based prediction for in-hospital pressure injury development using electronic health records: a retrospective observational cohort study in a university hospital in Japan. Int J Nurs Stud 2021 Jul;119:103932. [doi: 10.1016/j.ijnurstu.2021.103932] [Medline: 33975074]
38. Delparte JJ, Flett HM, Scovil CY, Burns AS. Development of the spinal cord injury pressure sore onset risk screening (SCI-Presors) instrument: a pressure injury risk decision tree for spinal cord injury rehabilitation. Spinal Cord 2021 Feb;59(2):123-131. [doi: 10.1038/s41393-020-0510-y] [Medline: 32694750]
39. Vyas K, Samadani A, Milosevic M, Ostadabbas S, Parvaneh S. Additional value of augmenting current subscales in Braden scale with advanced machine learning technique for pressure injury risk assessment. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE; 2020:2993-2995. [doi: 10.1109/BIBM49941.2020.9313401]
40. Sotoodeh M, Gero ZH, Zhang W, Hertzberg VS, Ho JC. Pressure ulcer injury in unstructured clinical notes: detection and interpretation. AMIA Annu Symp Proc 2021 Jan 25;2020:1160-1169. [Medline: 33936492]
41. Alderden J, Drake KP, Wilson A, Dimas J, Cummins MR, Yap TL. Hospital acquired pressure injury prediction in surgical critical care patients. BMC Med Inform Decis Mak 2021 Jan 6;21(1):12. [doi: 10.1186/s12911-020-01371-z] [Medline: 33407439]
42. Choi BK, Kim MS, Kim SH. Risk prediction models for the development of oral-mucosal pressure injuries in intubated patients in intensive care units: a prospective observational study. J Tissue Viability 2020 Nov;29(4):252-257. [doi: 10.1016/j.jtv.2020.06.002] [Medline: 32800513]
43. Hu YH, Lee YL, Kang MF, Lee PJ. Constructing inpatient pressure injury prediction models using machine learning techniques. Comput Inform Nurs 2020 Aug;38(8):415-423. [doi: 10.1097/CIN.0000000000000604] [Medline: 32205474]
44. 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 1;27(4):567-576. [doi: 10.1093/jamia/ocaa004] [Medline: 32065628]
45. Ahmad MA, Larson B, Overman S, et al. Machine learning approaches for pressure injury prediction. In: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI): IEEE; 2021:427-431. [doi: 10.1109/ICHI52183.2021.00069]
46. Walker SC, French B, Moore RP, et al. Model-guided decision-making for thromboprophylaxis and hospital-acquired thromboembolic events among hospitalized children and adolescents: the CLOT randomized clinical trial. JAMA Netw Open 2023 Oct 2;6(10):e2337789. [doi: 10.1001/jamanetworkopen.2023.37789] [Medline: 37831448]
47. Walker SC, Creech CB, Domenico HJ, French B, Byrne DW, Wheeler AP. A real-time risk-prediction model for pediatric venous thromboembolic events. Pediatrics 2021 Jun;147(6):e2020042325. [doi: 10.1542/peds.2020-042325] [Medline: 34011634]
48. 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]
49. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med 1989 May;8(5):551-561. [doi: 10.1002/sim.4780080504] [Medline: 2657958]
50. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat 1996 Jan;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
51. Assel M, Sjoberg DD, Vickers AJ. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Diagn Progn Res 2017 Dec;1:19. [doi: 10.1186/s41512-017-0020-3] [Medline: 31093548]
52. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med 2019 Sep 20;38(21):4051-4065. [doi: 10.1002/sim.8281] [Medline: 31270850]
53. Agarwal R, Domenico HJ, Balla SR, et al. Palliative care exposure relative to predicted risk of six-month mortality in hospitalized adults. J Pain Symptom Manage 2022 May;63(5):645-653. [doi: 10.1016/j.jpainsymman.2022.01.013] [Medline: 35081441]
54. Freundlich RE, Li G, Domenico HJ, Moore RP, Pandharipande PP, Byrne DW. A predictive model of reintubation after cardiac surgery using the electronic health record. Ann Thorac Surg 2022 Jun;113(6):2027-2035. [doi: 10.1016/j.athoracsur.2021.06.060] [Medline: 34329600]

55. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Eur Urol 2015 Jun;67(6):1142-1151. [doi: 10.1016/j.eururo.2014.11.025] [Medline: 25572824]
56. Domenico H, Reese T, Moore R, Byrne D, Hernandez T. Predicted risk of hospital acquired pressure injury calculator. Vanderbilt University Medical Center. 2023. URL: https://cqs.app.vumc.org/shiny/PressureInjuryPrediction/ [accessed 2023-08-13]
57. Byrne DW. Artificial Intelligence for Improved Patient Outcomes: Principles for Moving Forward with Rigorous Science: Lippincott Williams & Wilkins; 2022.
58. Reese TJ, Liu S, Steitz B, et al. Conceptualizing clinical decision support as complex interventions: a meta-analysis of comparative effectiveness trials. J Am Med Inform Assoc 2022 Sep 12;29(10):1744-1756. [doi: 10.1093/jamia/ocac089] [Medline: 35652167]
59. Weaver CGW, McAlister FA. Machine learning, predictive analytics, and the emperor's new clothes: why artificial intelligence has not yet replaced conventional approaches. Can J Cardiol 2021 Aug;37(8):1156-1158. [doi: 10.1016/j.cjca.2021.03.003] [Medline: 33711476]
60. Reese TJ, Mixon AS, Matheny ME, et al. Using intervention mapping to design and implement a multicomponent intervention to improve antibiotic and NSAID prescribing. Transl Behav Med 2023 Dec 15;13(12):928-943. [doi: 10.1093/tbm/ibad063] [Medline: 37857368]
61. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health Affairs 2014 Jul;33(7):1148-1154. [doi: 10.1377/hlthaff.2014.0352] [Medline: 25006140]
62. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar;28(3):231-237. [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
63. Gomez Lumbreras A, Reese TJ, Del Fiol G, et al. Shared decision-making for drug-drug interactions: formative evaluation of an anticoagulant drug interaction. JMIR Form Res 2022 Oct 19;6(10):e40018. [doi: 10.2196/40018] [Medline: 36260377]
64. Reese TJ, Del Fiol G, Morgan K, et al. A shared decision-making tool for drug interactions between warfarin and nonsteroidal anti-inflammatory drugs: design and usability study. JMIR Hum Factors 2021 Oct 26;8(4):e28618. [doi: 10.2196/28618] [Medline: 34698649]
65. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36. [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
66. Reese TJ, Schlechter CR, Kramer H, et al. Implementing lung cancer screening in primary care: needs assessment and implementation strategy design. Transl Behav Med 2022 Feb 16;12(2):187-197. [doi: 10.1093/tbm/ibab115] [Medline: 34424342]
67. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
68. Cabitza F, Zeitoun JD. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Ann Transl Med 2019 Apr;7(8):161. [doi: 10.21037/atm.2019.04.07] [Medline: 31168442]
69. Semler MW, Self WH, Wanderer JP, et al. Balanced crystalloids versus saline in critically ill adults. N Engl J Med 2018 Mar 1;378(9):829-839. [doi: 10.1056/NEJMoa1711584] [Medline: 29485925]

Abbreviations

AUC: area under the receiver operating curve
EHR: electronic health record
HAPI: hospital-acquired pressure injury
LASSO: least absolute shrinkage and selection operator
OR: odds ratio
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 14.08.23; peer-reviewed by J Walsh, L Zhang, N Fareed, P Okoro, P Kukhareva, W Wei; revised version received 08.03.24; accepted 10.03.24; published 08.05.24.

Please cite as:
Reese TJ, Domenico HJ, Hernandez A, Byrne DW, Moore RP, Williams JB, Douthit BJ, Russo E, McCoy AB, Ivory CH, Steitz BD, Wright A
Implementable Prediction of Pressure Injuries in Hospitalized Adults: Model Development and Validation
JMIR Med Inform 2024;12:e51842
URL: https://medinform.jmir.org/2024/1/e51842
doi:10.2196/51842

A Machine Learning Application to Classify Patients at Differing Levels of Risk of Opioid Use Disorder: Clinician-Based Validation Study

Tewodros Eguale[1,2,*], MD, PhD; François Bastardot[3,4,*], MSACI, MD; Wenyu Song[2,5], PhD; Daniel Motta-Calderon[6], MD; Yasmin Elsobky[2,7], BPharmSci, MSc; Angela Rui[2], MA; Marlika Marceau[8], BA; Clark Davis[2], BA; Sandya Ganesan[2], BA; Ava Alsubai[2], MA; Michele Matthews[1,9], PharmD; Lynn A Volk[8], MHS; David W Bates[2,5,8,10], MSc, MD; Ronen Rozenblum[2,5], MPH, PhD

1

2

3

4

5

6

7

8

9

10

[*]these authors contributed equally

Corresponding Author:
Ronen Rozenblum, MPH, PhD

Abstract

Background: Despite restrictive opioid management guidelines, opioid use disorder (OUD) remains a major public health concern. Machine learning (ML) offers a promising avenue for identifying and alerting clinicians about OUD, thus supporting better clinical decision-making regarding treatment.

Objective: This study aimed to assess the clinical validity of an ML application designed to identify and alert clinicians of different levels of OUD risk by comparing it to a structured review of medical records by clinicians.

Methods: The ML application generated OUD risk alerts on outpatient data for 649,504 patients from 2 medical centers between 2010 and 2013. A random sample of 60 patients was selected from 3 OUD risk level categories (n=180). An OUD risk classification scheme and standardized data extraction tool were developed to evaluate the validity of the alerts. Clinicians independently conducted a systematic and structured review of medical records and reached a consensus on a patient's OUD risk level, which was then compared to the ML application's risk assignments.

Results: A total of 78,587 patients without cancer with at least 1 opioid prescription were identified as follows: not high risk (n=50,405, 64.1%), high risk (n=16,636, 21.2%), and suspected OUD or OUD (n=11,546, 14.7%). The sample of 180 patients was representative of the total population in terms of age, sex, and race. The interrater reliability between the ML application and clinicians had a weighted kappa coefficient of 0.62 (95% CI 0.53-0.71), indicating good agreement. Combining the high risk and suspected OUD or OUD categories and using the review of medical records as a gold standard, the ML application had a corrected sensitivity of 56.6% (95% CI 48.7%-64.5%) and a corrected specificity of 94.2% (95% CI 90.3%-98.1%). The positive and negative predictive values were 93.3% (95% CI 88.2%-96.3%) and 60.0% (95% CI 50.4%-68.9%), respectively. Key themes for disagreements between the ML application and clinician reviews were identified.

Conclusions: A systematic comparison was conducted between an ML application and clinicians for identifying OUD risk. The ML application generated clinically valid and useful alerts about patients' different OUD risk levels. ML applications hold promise for identifying patients at differing levels of OUD risk and will likely complement traditional rule-based approaches to generating alerts about opioid safety issues.

(JMIR Med Inform 2024;12:e53625) doi:10.2196/53625

KEYWORDS

opioid-related disorders; opioid use disorder; machine learning; artificial intelligence; electronic health record; clinical decision support; model validation; patient medication safety; medication safety; clinical decision; decision making; decision support; patient safety; opioid use; drug use; opioid safety; medication; OUD; EHR; AI

Introduction

In the past few decades, the "opioid epidemic" has become a public health crisis. According to a 2020 US survey, 2.7 million people aged 12 years or older had an opioid use disorder (OUD), and only 1 in 9 (11.2%) received medication-assisted therapy [1]. OUD is a frequently underdiagnosed condition, and it is estimated that for every patient with an OUD diagnosis, there are at least 2 who remain undiagnosed [2]. In 2021, nearly 92,000 drug overdose deaths were reported in the United States [3]. Furthermore, 54% and 46% of the US $1.02 trillion aggregate annual societal costs in 2020 in the United States were attributed to overdose deaths and OUD, respectively [4].

There is an immediate urgency to identify patients at high risk of OUD and those with OUD. Clinicians have reported major barriers to adequately assessing patients' risk, including time pressure, incomplete or restricted medical records, and a lack of robust clinical decision support systems (CDSSs) [5,6]. The current rule-based approaches, such as Medicare Part D's Overutilization Monitoring System or statewide Prescription Drug Monitoring Programs, fail to incorporate clinical data and are often underused [7]. Moreover, unless CDSSs use individual patient-specific clinical data in generating alerts, many false positive alerts may be presented to clinicians contributing to alert fatigue [8].

Artificial intelligence and machine learning (ML) algorithms have recently demonstrated their usefulness in CDSSs; however, compared with conventional statistical methods, their black-box nature and a lack of studies assessing the clinical validity of these interventions have created uneasiness in the medical community [9-12]. MedAware is a commercial software application that uses various statistical and ML methods to identify and prevent medication safety issues, including the risk of OUD [13]. It uses an iterative development process and has conducted pilot testing to optimize its OUD risk prediction algorithm to increase its accuracy in patient risk identification.

The goals of this study were to assess the clinical validity of the ML application by (1) determining the agreement between the ML algorithm's output and the outcomes of structured clinicians' review of medical records in classifying patients into distinct categories of OUD risk, including not high risk, high risk, or suspected OUD or OUD; (2) determining the potential utility of using the ML application as an alerting tool by evaluating its test characteristics against the gold standard; and (3) identifying major factors contributing to discrepancies between the ML application and clinician risk assignments to provide a knowledge base for future system improvement.

Methods

Ethical Considerations

This study was approved by the Mass General Brigham Institutional Review Boards (#2014P002167) that granted a patient waiver of consent for this study. Patients did not receive any compensation.

Evaluation of the ML Application

MedAware (Ra'anana, Israel) has developed an ML software application to identify prescription errors and adverse drug events [13]. This application identifies medication issues based on ML methods including random forest algorithms—a widely used ML method in medical applications [14]. Multiple studies using ML models for disease prediction have achieved robust performance [15,16].

Based on clinical data in the electronic health record (EHR), the ML application's algorithms generate patient-specific alerts on medication orders that deviate from predominant prescribing patterns in similar patient situations. Previously, it was found that the ML application generates medication error alerts that might otherwise be missed with existing applications with a high degree of alert usefulness, and it has the potential to reduce costs [17,18].

The ML application has been enhanced to generate alerts in real time to identify patients at risk of OUD and overdose based on clinical, psychosocial, and medication data. The input features used in the model were age, gender, opioid and nonopioid medication history (for each prescription: drug name, route of administration, duration, and dosage), and diagnosis history found in *ICD-9* (*International Classification of Diseases*) diagnoses codes and problem lists. The application can also produce aggregate alert data about the risk of OUD or overdose, which may be used for population health management.

The model outcome was defined by MedAware by combining OUD diagnosis codes, medication use, and experts' annotation. The test cohort was independent from the training set to avoid overfitting. Random data splitting was conducted to separate training (50%) and test (50%) sets. MedAware used a scikit-learn (1.2.0) implementation of the random forest algorithm. It was used in a cross-fold manner and some of its hyperparameters (mainly: n_estimators, max_depth, class_weight) were tuned for optimization while leaving others at their default values. Additional details of the ML algorithm were not available to the research team because of intellectual property protections and were not the focus of this study; our study aimed to clinically validate OUD alerts generated by the algorithm against clinician judgement.

Study Setting and Patient Population

The patient population of this study comprised patients who had at least 1 outpatient encounter between January 1, 2012,

and December 31, 2013, and were prescribed at least 1 opioid medication between January 1, 2010, and December 31, 2013, in an outpatient setting at 2 large academic medical centers in the United States. Patients diagnosed with cancer and those with incomplete data were excluded. Once a patient had a documented OUD diagnosis or started receiving opioid rehabilitation drugs (eg, suboxone, naltrexone, methadone, and bezitramide), any subsequent patient data were excluded from the analysis as the patient's status was known.

The evaluated application classified patients into 3 levels of OUD risk: not high risk, high risk, and suspected OUD or OUD. Alerts to clinicians are generated for only the high risk and suspected OUD or OUD categories. The risk alerts are generated when a clinician initiates an opioid medication prescription. A short textual description is created by the application for each alert generated to explain why the alert fired, for example, *the patient has a long opioid sequence, concurrent benzodiazepines use*. This explanation enables clinicians to understand the general reasoning underlying the alert. To improve study efficiency, the validation study comprised a random sample of 60 patients from each risk category for a total of 180 cases for which a retrospective review was performed by clinicians [19].

Data Collection and Transfer

Clinical and encounter data on the patient population from 2010 to 2013 were extracted and sent to MedAware, including demographics, diagnoses, problem lists, outpatient and inpatient encounters, encounter clinicians, clinician specialties, procedures, medications, allergies, vital signs, and selected blood test outcomes. Patient and clinician names and medical record numbers were removed from the data set, and a random study ID was assigned to each patient and clinician before the limited data set was sent through a secure transfer application (password-protected and encrypted) for analysis.

Development of a Risk Classification Scheme and Pilot Testing

Evaluation criteria for risk assignment by clinicians using the clinical data were developed with an extensive review of established guidelines, such as those of the Centers for Disease Control and Prevention and *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*), and risk factors for OUD through an iterative process in consultation with experts in the field of pain and opioid management [20-24]. The research clinicians and team reviewed the Centers for Disease Control and Prevention's and *DSM-5* guidelines and created draft criteria based on these guidelines to reflect 3 levels of risk, and then these criteria were reviewed by 2 pain management experts (a physician and a pharmacist). After modifications, this risk classification scheme was piloted to evaluate its effectiveness and compatibility with the ML application. We conducted the pilot review of medical records with 25 randomly selected medical records. One research assistant (CD) extracted data from the medical records using a standardized data collection tool as described below and 2 physician reviewers (FB and TE) individually reviewed the data. The reviewers reached a consensus on their risk determinations, and revisions were made to criteria, as needed, to standardize assessments and support a more transparent, generalizable validation process. MedAware sent a list of those patients for whom a risk assessment was conducted to be used for selecting the random sample for review of medical records.

Structured Clinicians' Review of Medical Records Using a Standardized Data Collection Tool

In total, 180 patients with a history of opioid use were randomly selected from those patients classified by the ML application into 3 risk categories (60 in each group), and structured reviews of medical records were conducted to evaluate patients' OUD risk. Clinicians were blinded to the patients' risk assignment by the application. A data abstraction tool was developed to organize relevant patients' clinical data from an EHR and facilitate the process for the review of medical records (Figure 1). This tool contains important demographic, patient, and family medical history including psychiatric and psychosocial information, patient complaints as documented in relevant clinical notes, relevant laboratory findings and drug history with graphical representation of opioid drug start and stop dates (ie, medication timeline; Figure 2), clinical events relevant to pain management such as surgeries or dates of major accidents, admission and emergency room visits, and curated clinical notes related to relevant clinical events. Collected data included both structured and free-text data that were extracted by research staff and organized into the abstraction tool. Data collection was focused on relevant information during the 2010-2013 time period; however, as the complete medical record was available for review, relevant information available prior to 2010 may have been considered. After training, 5 research assistants (CD, AA, SG, AR, and MM) individually extracted clinical data. Information from medical records was reviewed by extractors and clinicians up to the ML application's first alert date (index date). For patients determined to be not high risk by the ML application, a random date was assigned up to which medical record data were extracted and reviewed.

Figure 1. Tool used to extract data from patients' medical records. Template used to organize patient information extracted during the review of electronic health records (EHRs). A patient's demographics and relevant past medical, psychosocial, family, and medication histories were captured. Provider notes and encounters relevant to opioid use and pain management were described and recorded by date. Any patient's laboratory findings relevant to opioid use or other medications of interest were also recorded. Clinician reviewers recorded their risk categorization and rationalization after reviewing the information captured on the data extraction tool and reviewing the EHR, as needed. OUD: opioid use disorder; PRN: pro re nata; SUD: substance use disorder.

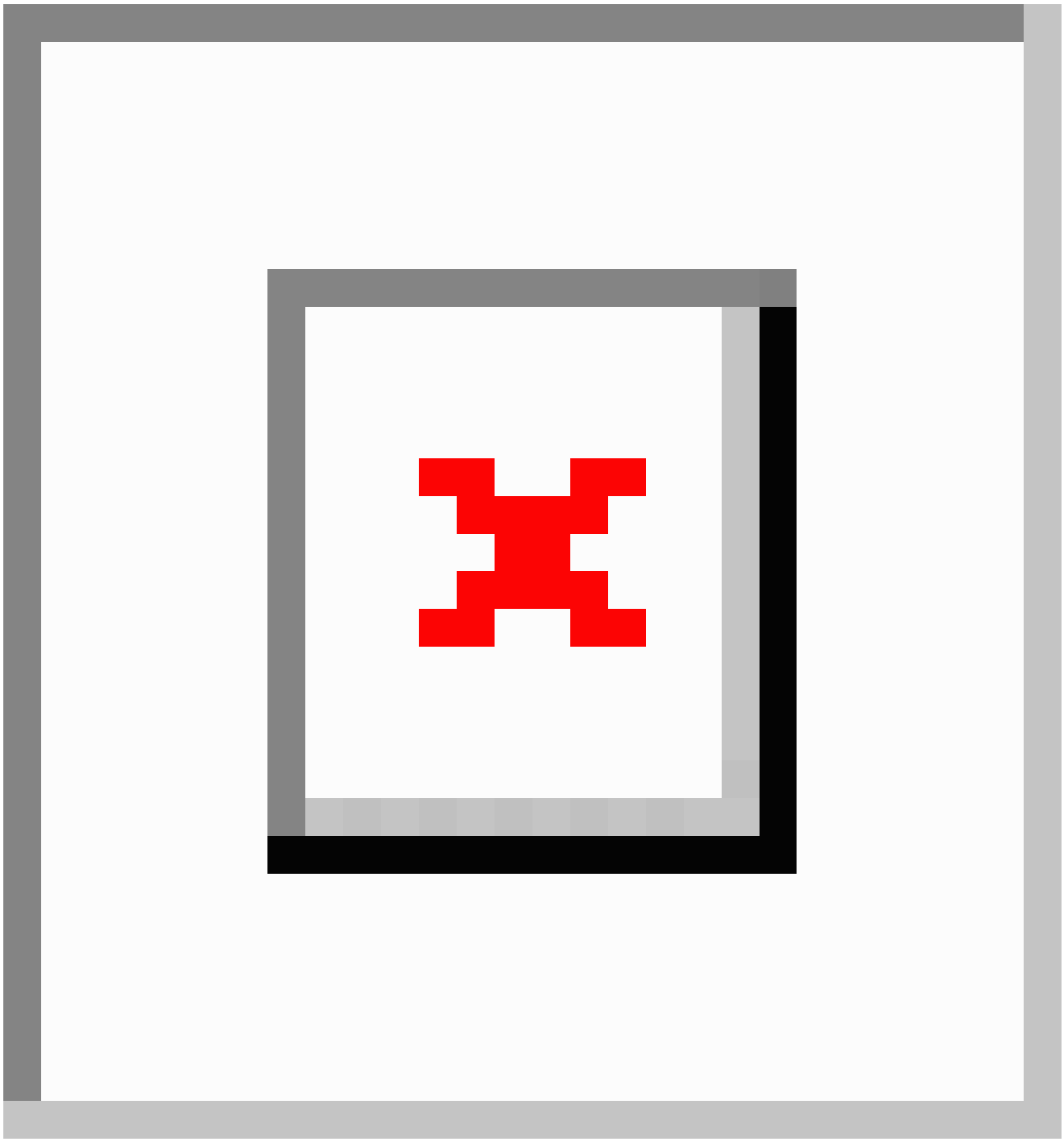
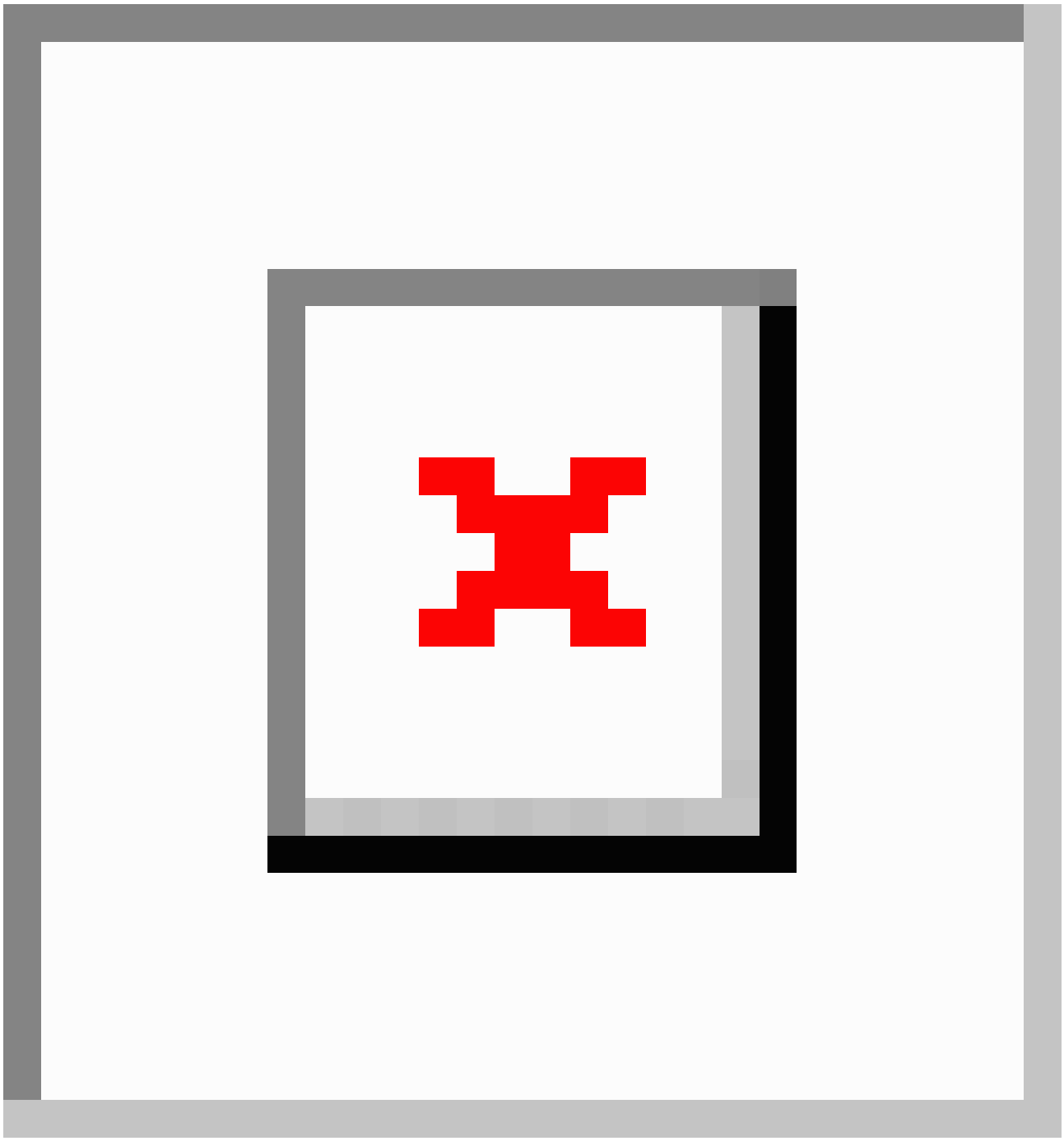


Figure 2. Example medication timeline from the tool used to extract data from the review of medical records. This timeline was created after medication history was recorded and morphine milligram equivalent (MME) conversion for opioid medications was included. Relevant encounters were recorded by date to provide context for the medication timeline (eg, surgery). Medications of interest included opioids, benzodiazepines, antidepressants, anticonvulsants, and other nonopioid medications contributing to risk. BID: twice a day; Q6H: every 6 hours; Q8H: every 8 hours; QD: every day, daily; SR: sustained release.



Four clinician reviewers (FB, TE, DM, and YE) individually examined data extracted by the research assistants and reviewed the EHRs directly, as needed, to holistically understand the clinical context of opiate prescription for the patients. The reviewers comprised 2 general internal medicine physicians, a hospitalist with extensive daily opioid prescribing experience, one with a PhD focusing on pharmacoepidemiology and drug safety, a recent medical student, and a pharmacist. All medical records were reviewed by 2 independent clinicians. After the primary review of medical records, a second reviewer blinded to the risk assignment of the first reviewer determined the risk level for OUD. The 2 clinicians discussed the case to reach consensus when their risk assignments differed. This consensus determination was then compared to the ML application's alert. Statistical analyses were conducted to evaluate the level of agreement between the clinician reviewers and the ML application's risk classifications.

Evaluation of Reasons for Disagreement Between Risk Assignments

To evaluate and identify the main reasons for disagreement between the clinician reviewers and the ML application's risk

classifications, a qualitative analysis was also conducted. For cases where there was disagreement, additional information contributing to the system risk assessment was requested from MedAware. Using a thematic analysis approach, 3 members of the research team (AR, LAV, and MM) independently conducted a qualitative analysis of the alert information. They reviewed the ML application's reasoning for assigning a particular risk category, information from the data extraction sheet, and information from the clinician reviewer's final risk assignment consensus. Then, this information was systematically coded to identify, categorize, and sort key concepts for the disagreements. Codes were then grouped into emergent themes and relationships after iterative review and discussion. In cases where there was disagreement, all 3 researchers reviewed and discussed the case together to reach consensus.

Statistical Analysis

We used descriptive statistics to summarize demographic characteristics of the study population, patients in each of the 3 risk categories identified by the ML application, and the 180 patients sampled for the validation study. We assessed the validity of the application by comparing them to the structured clinicians' review of medical records. The agreements between the 2 methods were evaluated with the following parameters:

1. Overall percent agreements were calculated, including percent agreements for the 3 risk categories. Disagreements were reported for the overall validated sample and the 3 opioid risk categories.
2. Weighted kappa and 95% CIs were reported because of the ordered nature of the risk categories to measure the agreement between the 2 methods.
3. Naïve sensitivity and naïve specificity were calculated along with positive and negative predictive values for the ML application using the structured clinicians' review of medical records as a gold standard and combining the 2 opioid risk categories, namely high risk and suspected OUD or OUD.
4. Corrected sensitivity and corrected specificity were calculated to account for verification bias, that is, overestimation of sensitivity and underestimation of specificity [19,25,26]. Verification bias occurs when disease status (eg, the presence or absence of OUD) is not ascertained in all participants by the gold-standard method (review of medical records) and proportionately more high risk and suspected OUD or OUD patients identified by the test methodology (eg, the ML algorithm) were selected for verification. This verification-biased sampling increases sensitivity and decreases specificity, and these parameters are mathematically corrected to adjust for the biased sampling method.
5. Descriptive statistics were calculated for evaluating risk assignments to determine the most frequently occurring themes for disagreement between the 2 methods.

Results

Patient Risk Categories and Demographics

Of the 649,504 eligible patients with at least 1 prescription in the source data, 78,587 (12.1%) were classified by the ML application into the 3 risk categories after excluding patients with no opioid prescription, patients without sufficient data to evaluate opioid risk, or patients with a diagnosis of cancer (Figure 3). Patients were excluded due to insufficient data if they did not have 1 day before and 1 year of data after their first opioid prescription, or if they were identified as having OUD (based on a diagnosis or rehabilitation drug) and did not have a first opioid prescription before identification of OUD. Patients with opioids prescribed within 2 years of a cancer diagnosis based on *ICD-9* (*International Classification of Diseases, Ninth Revision*) codes were excluded. Accordingly, 50,405 (64.1%) patients were classified by the ML application as being in the not high risk category, 16,636 (21.2%) as being in the high risk category, and 11,546 (14.7%) as being in the suspected OUD or OUD category. We excluded patients who do not have 1 day before and 1 year of data after the first opioid Rx or, if identified as having OUD (based on diagnosis or rehabilitation drug) and do not have a first opioid Rx before identification.

Table 1 details the distribution of eligible patients by demographic characteristics across the different ML application risk assignment categories and sampled patients. Female sex and age 30-64 years were overrepresented in the groups with opioid prescriptions and validation samples for medical records review compared to the eligible patient pool. The sample randomly selected for validation with the structured review of medical records was representative of the patients on opioid treatment with regard to age, sex, and race.

Figure 3. Patient flow diagram with the final verification sample. Patients were excluded from the overall population if they did not have any opioid prescriptions since 2010, were diagnosed with cancer, or had insufficient data to predict opioid risk. The remaining patients were evaluated for opioid risk and stratified by risk classification category. A total of 60 patients were randomly sampled from each risk classification category to be used for the review of medical records and clinician evaluation.

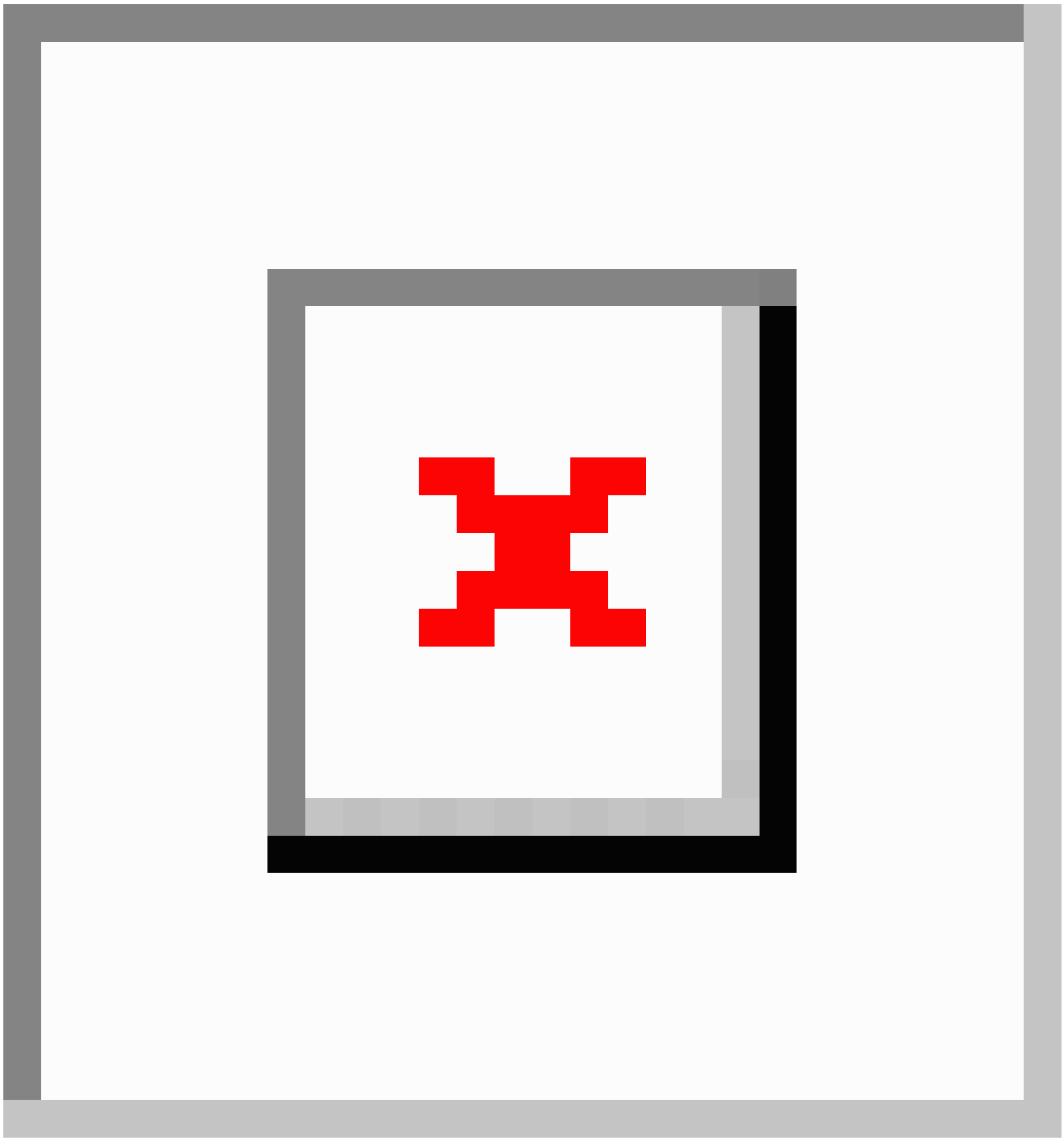


Table . Demographic characteristics of the study populations, including the overall patient population, patients who met the criteria for opioid risk evaluation, patients stratified by the machine learning (ML) application's risk categories "not high risk," "high risk," or "suspected OUD or OUD," and for the validation sample for total review of medical records.

Patient characteristics		Patients with at least 1 prescription (n=649,504), n (%)	Patients meeting criteria for opioid risk evaluation (n=78,587), n (%)	Not high risk (n=50,405), n (%)	High risk (n=16,636), n (%)	Suspected OUD or OUD (n=11,546), n (%)	Sampled patients (n=180), n (%)
Sex							
	Female	385,959 (59.4)	50,064 (63.7)	31,184 (61.9)	11,860 (71.3)	7020 (60.8)	116 (64.4)
	Male	263,535 (40.6)	28,521 (36.3)	19,221 (38.1)	4775 (28.7)	4525 (39.2)	64 (35.6)
	Unknown	10 (0.0)	2 (0.0)	0 (0.0)	1 (0.0)	1 (0.0)	0 (0.0)
Age (years)							
	0 - 17	76,024 (11.7)	737 (0.9)	635 (1.3)	77 (0.5)	25 (0.2)	4 (2.2)
	18 - 29	83,216 (12.8)	8467 (10.8)	5953 (11.8)	1645 (9.9)	869 (7.5)	22 (12.2)
	30 - 49	180,603 (27.8)	27,666 (35.2)	17,777 (35.3)	5834 (35.1)	4055 (35.1)	64 (35.6)
	50 - 64	164,188 (25.3)	24,329 (31.0)	14,442 (28.7)	5564 (33.4)	4323 (37.4)	51 (28.3)
	≥65	145,473 (22.4)	17,388 (22.1)	11,598 (23.0)	3516 (21.1)	2274 (19.7)	39 (21.7)
Race [a]							
	American Indian or Native Alaskan	719 (0.1)	89 (0.1)	60 (0.1)	16 (0.1)	13 (0.1)	0 (0)
	Asian	28,328 (4.4)	2211 (2.8)	1763 (3.5)	297 (1.8)	151 (1.3)	3 (1.7)
	Black or African American	41,794 (6.4)	6189 (7.9)	4119 (8.2)	1245 (7.5)	825 (7.1)	11 (6.1)
	Native Hawaiian or Pacific Islander	286 (0.0)	26 (0.0)	20 (0.0)	5 (0.0)	1 (0.0)	0 (0)
	White	475,939 (73.3)	58,617 (74.6)	36,972 (73.3)	12,373 (74.4)	9272 (80.3)	135 (75.0)
	Other (unknown, declined, or bi- or multiracial)	102,438 (15.8)	11,455 (14.6)	7471 (14.8)	2700 (16.2)	1284 (11.1)	31 (17.2)
Ethnicity [a]							
	Hispanic or Latino	43,119 (6.6)	1874 (2.4)	1279 (2.5)	420 (2.5)	175 (1.5)	17 (9.4)
	Other[b]	606,385 (93.4)	76,713 (97.6)	49,126 (97.5)	16,216 (97.5)	11,371 (98.5)	163 (90.6)

[a]Race and ethnicity data are based on coded fields in the electronic health record.

[b]Other refers to non-Hispanic or non-Latino, declined to respond, and unknown.

Percent Agreement and Kappa Statistics

Prior to conducting final consensus assessments, the independent clinician reviewers' assessment of the levels of risk matched exactly for 70% (126/180) of the patients. When comparing assessments of the not high risk group and those of the high risk and suspected OUD or OUD groups, the clinician reviewer assessments matched 88% of the time.

The overall percent agreement between the ML application and clinician reviewers in stratifying patients into 3 risk categories was 70% (126/180 patients; Table 2). Of the 30% disagreements, 22.8% (n=41) and 7.2% (n=13) indicated underestimation and overestimation of risk by the ML application, respectively, compared to the clinicians' structured review of medical records. Among different risk categories, percent agreement was the highest (90%) for the suspected OUD or OUD category than for the not high risk and high risk categories (60% each). Of the patients classified to the suspected OUD or OUD category by the ML application, 8.3% and 1.7% of them were classified to the high risk and not high risk categories, respectively, by the clinicians' review of medical records. Of the patients classified to the not high risk category by the ML application, clinician reviews classified 40% of patients to the 2 higher risk categories: 30% of patients to the high risk category and 10% of patients to the suspected OUD or OUD category.

Table . Distribution in opioid risk assignment between the machine learning (ML) application and clinicians' structured review of medical records of 180 randomly sampled patients (percent agreement 70%, 95% CI 63.3%-76.7%; weighted kappa coefficient 0.62, 95% CI 0.52-0.71).

ML system risk assignment	Clinician reviewer risk assignment, n			
	Not high risk	High risk	Suspected OUD/OUD	Total
Not high risk	36	18	6	60
High risk	7	36	17	60
Suspected OUD[a] or OUD	1	5	54	60

[a]OUD: opioid use disorder.

The interrater reliability, as expressed using the weighted kappa coefficient for the 2 methods, was 0.62 (95% CI 0.53-0.71), indicating good or substantial agreement [27].

Corrected Sensitivity, Corrected Specificity, and Positive and Negative Predictive Values

Table 3 presents a revised version of Table 2, where the 2 higher-level opioid risk categories (high risk and suspected OUD or OUD) were combined to investigate the potential utility of the ML application in generating signals or alerts to prescribing clinicians, that is, how complete and accurate the ML application is in identifying patients who are at the risk of developing or who may already have OUD. The naïve sensitivity of the ML application was 82.4% (95% CI 75.9%-88.9%), and its naïve specificity was 81.8% (95% CI 70.2%-93.4%). After accounting for verification-biased sampling, the corrected sensitivity of the ML application was 56.6% (95% CI 48.7%-64.5%) and its corrected specificity was 94.2% (95% CI 90.3%-98.1%). The positive and negative predictive values of the ML application were 93.3% (95% CI 88.2%-96.3%) and 60.0% (95% CI 50.4%-68.9%), respectively.

Table . Distribution in opioid use disorder (OUD) risk assignment between the machine learning (ML) application and clinicians' structured review of medical records when the 2 higher-risk categories were combined to investigate the utility of an OUD risk alert at the time of prescribing.

ML system risk assignment	Clinician reviewer risk assignment, n		
	High risk and suspected OUD or OUD	Not high risk	Total
High risk and suspected OUD or OUD	112	8	120
Not high risk	24	36	60
Total	136	44	180

Key Reasons for Disagreements in OUD Risk Categories Between the ML Application and Clinician Reviewers

Table 4 contains the 6 themes that emerged as reasons for disagreements between the ML application and the clinicians' structured review of medical records after conducting a qualitative analysis. Disagreement between the 2 methods was noted for 54 patients, among whom the ML application underestimated the OUD risk in 41 patients and overestimated it in 13 patients. Two or more themes were identified as reasons for most of the disagreements (74.9%). Of the 6 themes, the theme "differences in risk assessment of medication information," accounted for most of the disagreements (72%), followed by the theme "information in clinical notes not available to the ML application" (55.6%).

Table . Key reasons for disagreements in opioid use disorder (OUD) risk assignments between the machine learning (ML) application and clinician reviewers. The reasons for discrepancies were categorized into 6 major themes. More than 1 reason might be identified for a given patient. Results are displayed by whether the assigned risk category was underestimated or overestimated by the ML application in comparison with the clinician reviewers.

Themes of reasons for disagreements in OUD risk assignment	Description of the themes	Patients with at least 1 reason coded in a given theme category, n (%)		
		Cases underestimated by MedAware[a] (n=41)	Cases overestimated by MedAware[b] (n=13)	Total discrepant cases (n=54)
I. Differences in risk assessment of medication information	Medication information available to both the clinician reviewers and the MedAware system contributed to differing risk assessments (eg, medication duration, dose, indication, and gaps in medication timelines).	30 (73.2)	9 (69.2)	39 (72.2)
II. Information in clinical notes not available to MedAware system	Information in patients' clinical notes was available to the clinician reviewers but not to the MedAware system (eg, psychosocial information, experience with opioids and other medications, patient participation in pain management and substance abuse services, and medication information not on the medication list).	27 (65.9)	3 (23.1)	30 (55.6)
III. Differences in risk assessment of psychosocial issues	Psychosocial or psychiatric information available to both the clinician reviewers and the MedAware system contributed to differing risk assessments (eg, patient history of substance abuse, family members with a history of psychosocial or psychiatric issues, and the presence of patients' individual psychiatric conditions contributed to differing risk assessments).	17 (41.5)	2 (15.4)	19 (35.2)
IV. Differences in risk assessment of nonopioid medications	Information on nonopioid medications available to both research reviewers and the MedAware system, which reflects an increased complexity of the patient's medical situation (eg, pain level) or a higher risk when combined with opioids, contributed to differences in risk assessments (eg, zolpidem and gabapentinoids).	10 (24.4)	2 (15.4)	12 (22.2)
V. Bugs identified in the MedAware system	Bugs in the MedAware system included inaccurate mapping of data elements (eg, dosage units and incorrect medication), missing medication in drug class, and incorrectly constructed alert messages.	5 (12.2)	5 (38.5)	10 (18.5)

Themes of reasons for disagreements in OUD risk assignment	Description of the themes	Patients with at least 1 reason coded in a given theme category, n (%)		
		Cases underestimated by MedAware[a] (n=41)	Cases overestimated by MedAware[b] (n=13)	Total discrepant cases (n=54)
VI. Presence of other clinical information not considered by the MedAware system or the clinician reviewers	Clinical information that may indicate the risk of OUD not considered by the clinician reviewers or the MedAware system, but not both, such as hepatitis C diagnosis, urine toxicity tests, and MedAware system access to *ICD-9*[c] diagnostic information that clinician reviewers did not see.	6 (14.6)	0 (0.0)	6 (11.1)

[a]The ML application's risk assignment was lower in severity compared to the clinician reviewers' risk assignment.

[b]The ML application's risk assignment was higher in severity compared to the clinician reviewers' risk assignment.

[c]*ICD-9*: *International Classification of Diseases, Ninth Revision*.

Discussion

Principal Results

ML algorithms can leverage large-scale EHR and medical claims data and potentially identify patients at risk of OUD [28-32]. However, very few studies have assessed the clinical validity and potential utility of ML algorithms designed to differentiate among levels of patients' OUD risk. In this study, we examined the agreement between an ML application and clinicians' structured review of medical records in classifying patients on opioid drug treatment into 3 distinct categories of OUD risk (ie, not high risk, high risk, or suspected OUD or OUD). We also assessed the application's utility in identifying clinically valid alerts and identified and quantified reasons that could lead to disagreements between clinicians' judgment and outputs of ML applications. The ML application was validated in an outpatient database, and it appeared to have value.

There was substantial agreement between the application and the clinician reviewers' structured review of medical records. The agreement between the 2 methods was the highest for the suspected OUD or OUD category. The ML application correctly identified this most vulnerable group of patients to increase clinician awareness and responsiveness to improve patient management, including modifications to their medication regimen or referral to a specialized treatment service to mitigate the complications of opioid use. Moreover, if the ML application is used to generate alerts on patients at high risk of OUD or those who already have OUD, it will identify approximately 60% of these patients with a 93.3% precision (positive predictive value). Thus, the results of this study show that this ML application was able to generate clinically valid and useful alerts to screen for patients at risk of OUD. It is important to recognize that alerting clinicians regarding patients at risk of OUD should be coupled with clinician education on appropriate treatment guidelines and practices to avoid undertreatment of pain and patient stigma [33,34].

Comparison With Prior Work

Previous studies have shown that artificial intelligence tools using ML algorithms can improve treatment, enhance quality of care and patient safety, reduce burden on providers, and generally increase the efficiency with which resources are used, resulting in potential cost savings or health gains [7,32,35-38]. In addition, our findings align with those of previous studies that highlight the potential of ML applications to predict individual patients' risk of specific medical conditions and associated complications to offer specialized care programs to high-risk patients [39,40]. Our study also confirms and extends the findings of a few studies that examined other ML applications and highlighted the potential to identify patients at risk for substance misuse and abuse, including OUD and opioid overdose [31,38,41]. Nevertheless, these comparable ML applications were plagued with very low positive predictive values due, in part, to low OUD prevalence as a result of suboptimal definitions of OUD by relying solely on *ICD* (*International Classification of Diseases*) codes [42]. A few previous studies identified additional limitations and challenges related to comparable ML applications. For example, Afshar et al [43] described the use of an algorithm to identify patients at risk for any substance misuse at the time of admission, based on clinical notes from the first 24 hours after hospital admission. In this study, we found that the positive predictive value of this tool was 61%-72%, which was lower than that of the ML application. The tool that Afshar et al [43] studied does not identify patients outside of the hospital setting and depends on physicians' notes. As a result, this tool is not suited for more general screening using structured clinical EHR data and medical claims data. Another recent study by Lo-Ciganic et al [41] described an algorithm to predict the occurrence of overdose episodes, but does not identify patients who are most at risk of OUD in the future.

We believe that the substantial agreement, high specificity, and high positive predictive value of the ML application was achieved because we pilot-tested the ML models in comparison with clinician assessments and then used an iterative process

with continuous calibration of model parameters to optimize the accurate identification of OUD risk categories. In addition, we used a composite definition of OUD not restricted to *ICD* codes resulting in a higher prevalence of OUD identified in the patient population. The ML application classified 1 in 7 and about one-fifth of the eligible population with prescribed opioids in the suspected OUD or OUD and high risk categories, respectively, compared to other studies that reported a prevalence of OUD in the range of 1%-5% [44,45]. Furthermore, the full accessibility of the EHR at the time of case evaluation, coupled with standardized data extraction and a medication timeline visualization tool, allowed seamless analysis of cases contributing to the high accuracy rates.

Our study also identified the main reasons for disagreements between the clinician reviewers and the ML application's risk assignments. These reasons included information available in the clinical notes not being accessible to the ML application (eg, psychosocial issues and patients' participation in substance abuse services), and different interpretation of available information such as differences in the impact of antidepressant treatments. Clinicians considered stable and sufficiently treated depression as not being a risk factor for OUD [46]. In analyzing the reasons for discrepancies, we observed factors related to model training processes, data quality, and outcome definitions. The knowledge gained through our analytic process could be useful to further optimize their ML algorithm development pipeline. As of today, it is critical to standardize the ML development process and make it more understandable to clinical end users. However, to our knowledge, few efforts have been made to systematically analyze each component of the model development process from the clinician's point of view and further evaluate its impact on the model's clinical implementation. We believe that our work can facilitate a better bridging of the gap between ML model builders and clinicians.

Limitations

Our study has some limitations. We used retrospective data to evaluate an algorithm primarily designed to be used in real time. Although many of the findings from our retrospective analysis should be applicable to real-time alerting, it is difficult to predict whether some alerts would perform differently or how clinicians would respond to real-time alerts. Second, although our clinician reviewers were carefully trained and a coding manual was developed with clear operational definitions, each risk assessment required a degree of judgment on the part of the reviewers; human factors could impact the final risk assignment. Finally, our study was limited to outpatients at 2 large academic medical centers in the United States, which limits the generalizability of our results. Additional biases may have been introduced into the ML application in ways that the research team were not able to assess [7,47]. Although the total population of patients receiving outpatient care within an academic medical center was included, there may have been biases in patients who were able to access care, those receiving opioid prescriptions, and in the clinical documentation of concerns regarding opioid use and substance abuse. Validation across different sites and populations (eg, veterans' facilities) may reveal site-specific differences and may require unique models or warrant the identification and capture of new descriptive features.

Conclusions

We tested an ML application that assessed OUD risk in an extensive outpatient EHR database and found that it appeared to classify patients into differing levels of OUD risk, and that there was substantial agreement with clinicians' review of medical records. We identified key themes for disagreements between the commercial application and clinician review, which can be used to further enhance ML applications. ML algorithms applied to available EHR clinical data hold promise for identifying patients at differing levels of OUD risk and supporting better clinical decision-making regarding treatment. Such tools will likely complement traditional, rule-based approaches to provide alerts about potential opioid prescribing safety issues.

Acknowledgments

This work was supported in part by MedAware, Ltd. DWB reports grants and personal fees from EarlySense, personal fees from Center for Digital Innovation Negev, equity from ValeraHealth, equity from Clew, equity from MDClone, personal fees and equity from AI-Enhanced Safety of Prescription, and grants from Merative, outside the submitted work. RR reports having equity from Hospitech Respiration, equity from Tri.O Medical Device, equity from AEYE Health, equity from RxE2, equity from OtheReality; equity from Co-Patient Support, and equity from Medyx.ai, all of which are unrelated to this work. He is also receiving research funding from Telem, Calosense Health, Breath of Health, and BriefCam.

MedAware's contributions to this study were limited to running the patient data through their ML application, providing risk assessment results from their system, and additional information on selected patients to clarify their risk assessment results. They were not involved in any data analysis or interpretation. They had no influence on the results, and they were not part of the manuscript writing process.

Data Availability

The data sets generated in this study are not publicly available due to hospital institutional review board (IRB) regulations and patient privacy policies, but deidentified data sets are available from the corresponding author upon reasonable request. These deidentified data sets would include the model training set to facilitate independent model replications, patient demographics, and risk level assessments generated by the MedAware system and clinician review for the study cohort of 180 patients. Detailed

electronic medical record data extracted on this cohort of patients in support of clinician risk assessments will not be available due to IRB and institutional policy restricting the use of clinical notes and sharing of patient-sensitive data. The MedAware system algorithm will not be available for sharing as this is a proprietary commercial product.

Conflicts of Interest

None declared.

References

1. 2020 National Survey on Drug Use and Health (NSDUH): methodological summary and definitions. Substance Abuse and Mental Health Services Administration. 2021. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt35330/2020NSDUHMethodSummDefs091721.pdf [accessed 2023-07-25]
2. Kirson NY, Shei A, Rice JB, et al. The burden of undiagnosed opioid abuse among commercially insured individuals. Pain Med 2015 Jul;16(7):1325-1332. [doi: 10.1111/pme.12768] [Medline: 25929289]
3. Drug overdose death rates. National Institute on Drug Abuse. 2023. URL: https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates [accessed 2023-07-25]
4. Kuehn BM. Massive costs of the US opioid epidemic in lives and dollars. JAMA 2021 May 25;325(20):2040. [doi: 10.1001/jama.2021.7464]
5. Satterwhite S, Knight KR, Miaskowski C, et al. Sources and impact of time pressure on opioid management in the safety-net. J Am Board Fam Med 2019;32(3):375-382. [doi: 10.3122/jabfm.2019.03.180306] [Medline: 31068401]
6. Harle CA, Bauer SE, Hoang HQ, Cook RL, Hurley RW, Fillingim RB. Decision support for chronic pain care: how do primary care physicians decide when to prescribe opioids? A qualitative study. BMC Fam Pract 2015 Apr 14;16:48. [doi: 10.1186/s12875-015-0264-3] [Medline: 25884340]
7. Artificial intelligence in health care: benefits and challenges of technologies to augment patient care. United States Government Accountability Office. 2022. URL: https://www.gao.gov/assets/720/711471.pdf [accessed 2023-07-25]
8. 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. [Medline: 24940129]
9. Petersen C, Smith J, Freimuth RR, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. J Am Med Inform Assoc 2021 Mar 18;28(4):677-684. [doi: 10.1093/jamia/ocaa319] [Medline: 33447854]
10. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021 Nov;3(11):e745-e750. [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
11. Chen D, Liu S, Kingsbury P, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med 2019;2:43. [doi: 10.1038/s41746-019-0122-0] [Medline: 31304389]
12. Ghassemi M, Mohamed S. Machine learning and health need better values. NPJ Digit Med 2022 Apr 22;5(1):51. [doi: 10.1038/s41746-022-00595-9] [Medline: 35459793]
13. Technology: your safety layer within. MedAware. 2023. URL: https://www.medaware.com/technology/ [accessed 2022-11-04]
14. Syrowatka A, Song W, Amato MG, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: a scoping review. Lancet Digit Health 2022 Feb;4(2):e137-e148. [doi: 10.1016/S2589-7500(21)00229-6] [Medline: 34836823]
15. Hanko M, Grendár M, Snopko P, et al. Random forest-based prediction of outcome and mortality in patients with traumatic brain injury undergoing primary decompressive craniectomy. World Neurosurg 2021 Apr;148:e450-e458. [doi: 10.1016/j.wneu.2021.01.002] [Medline: 33444843]
16. Song W, Kang MJ, Zhang L, 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. [doi: 10.1093/jamia/ocaa336] [Medline: 33517452]
17. Rozenblum R, Rodriguez-Monguio R, Volk LA, et al. Using a machine learning system to identify and prevent medication prescribing errors: a clinical and cost analysis evaluation. Jt Comm J Qual Patient Saf 2020 Jan;46(1):3-10. [doi: 10.1016/j.jcjq.2019.09.008] [Medline: 31786147]
18. Schiff GD, Volk LA, Volodarskaya M, et al. Screening for medication errors using an outlier detection system. J Am Med Inform Assoc 2017 Mar 1;24(2):281-287. [doi: 10.1093/jamia/ocw171] [Medline: 28104826]
19. Irwig L, Glasziou PP, Berry G, Chock C, Mock P, Simpson JM. Efficient study designs to assess the accuracy of screening tests. Am J Epidemiol 1994 Oct 15;140(8):759-769. [doi: 10.1093/oxfordjournals.aje.a117323] [Medline: 7942777]
20. Dowell D, Haegerich TM, Chou R. CDC guideline for prescribing opioids for chronic pain—United States, 2016. MMWR Recomm Rep 2016 Mar 18;65(1):1-49. [doi: 10.15585/mmwr.rr6501e1] [Medline: 26987082]
21. American Psychiatric Association. Opioid use disorder. In: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition: American Psychiatric Association; 2013:541-546 URL: https://psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596 [accessed 2024-05-08] [doi: 10.1176/appi.books.9780890425596]
22. Webster LR. Risk factors for opioid-use disorder and overdose. Anesth Analg 2017 Nov;125(5):1741-1748. [doi: 10.1213/ANE.0000000000002496] [Medline: 29049118]

23. Burcher KM, Suprun A, Smith A. Risk factors for opioid use disorders in adult postsurgical patients. Cureus 2018 May 11;10(5):e2611. [doi: 10.7759/cureus.2611] [Medline: 30018867]
24. Zhao S, Chen F, Feng A, Han W, Zhang Y. Risk factors and prevention strategies for postoperative opioid abuse. Pain Res Manag 2019;2019:7490801. [doi: 10.1155/2019/7490801] [Medline: 31360271]
25. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983 Mar;39(1):207-215. [Medline: 6871349]
26. Punglia RS, D'Amico AV, Catalona WJ, Roehl KA, Kuntz KM. Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. N Engl J Med 2003 Jul 24;349(4):335-342. [doi: 10.1056/NEJMoa021659] [Medline: 12878740]
27. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363. [Medline: 15883903]
28. Burgess-Hull AJ, Brooks C, Epstein DH, Gandhi D, Oviedo E. Using machine learning to predict treatment adherence in patients on medication for opioid use disorder. J Addict Med 2023;17(1):28-34. [doi: 10.1097/ADM.0000000000001019] [Medline: 35914118]
29. Lo-Ciganic WH, Donohue JM, Yang Q, et al. Developing and validating a machine-learning algorithm to predict opioid overdose among medicaid beneficiaries in two US States: a prognostic modeling study. Lancet Digit Health 2022 Jun;4(6):e455-e465. [doi: 10.1016/S2589-7500(22)00062-0] [Medline: 35623798]
30. Afshar M, Sharma B, Dligach D, 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. [doi: 10.1016/S2589-7500(22)00041-3] [Medline: 35623797]
31. Heo KN, Lee JY, Ah YM. Development and validation of a risk-score model for opioid overdose using a national claims database. Sci Rep 2022 Mar 23;12(1):4974. [doi: 10.1038/s41598-022-09095-y] [Medline: 35322156]
32. Dong X, Deng J, Rashidian S, et al. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. J Am Med Inform Assoc 2021 Jul 30;28(8):1683-1693. [doi: 10.1093/jamia/ocab043] [Medline: 33930132]
33. Keister LA, Stecher C, Aronson B, McConnell W, Hustedt J, Moody JW. Provider bias in prescribing opioid analgesics: a study of electronic medical records at a hospital emergency department. BMC Public Health 2021 Aug 6;21(1):1518. [doi: 10.1186/s12889-021-11551-9] [Medline: 34362330]
34. Pergolizzi JV, Lequang JA, Passik S, Coluzzi F. Using opioid therapy for pain in clinically challenging situations: questions for clinicians. Minerva Anestesiol 2019 Aug;85(8):899-908. [doi: 10.23736/S0375-9393.19.13321-4] [Medline: 30871302]
35. Goecks J, Jalili V, Heiser LM, Gray JW. How machine learning will transform biomedicine. Cell 2020 Apr 2;181(1):92-101. [doi: 10.1016/j.cell.2020.03.022] [Medline: 32243801]
36. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. J Glob Health 2018 Dec;8(2):020303. [doi: 10.7189/jogh.08.020303] [Medline: 30405904]
37. Li Q, Wright J, Hales R, Voong R, McNutt T. A digital physician peer to automatically detect erroneous prescriptions in radiotherapy. NPJ Digit Med 2022 Oct 21;5(1):158. [doi: 10.1038/s41746-022-00703-9] [Medline: 36271138]
38. Canan C, Polinski JM, Alexander GC, Kowal MK, Brennan TA, Shrank WH. Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. J Am Med Inform Assoc 2017 Nov 1;24(6):1204-1210. [doi: 10.1093/jamia/ocx066] [Medline: 29016967]
39. Chen JH, Asch SM. Machine learning and prediction in medicine: beyond the peak of inflated expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509. [doi: 10.1056/NEJMp1702071] [Medline: 28657867]
40. Beaulieu-Jones BK, Yuan W, Brat GA, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med 2021 Mar 30;4(1):62. [doi: 10.1038/s41746-021-00426-3] [Medline: 33785839]
41. Lo-Ciganic WH, Huang JL, Zhang HH, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. JAMA Netw Open 2019 Mar 1;2(3):e190968. [doi: 10.1001/jamanetworkopen.2019.0968] [Medline: 30901048]
42. Lagisetty P, Garpestad C, Larkin A, et al. Identifying individuals with opioid use disorder: validity of International Classification of Diseases diagnostic codes for opioid use, dependence and abuse. Drug Alcohol Depend 2021 Apr 1;221:108583. [doi: 10.1016/j.drugalcdep.2021.108583] [Medline: 33662670]
43. Afshar M, Sharma B, Bhalla S, et al. External validation of an opioid misuse machine learning classifier in hospitalized adult patients. Addict Sci Clin Pract 2021 Mar 17;16(1):19. [doi: 10.1186/s13722-021-00229-7] [Medline: 33731210]
44. Barocas JA, White LF, Wang J, et al. Estimated prevalence of opioid use disorder in Massachusetts, 2011-2015: a capture-recapture analysis. Am J Public Health 2018 Dec;108(12):1675-1681. [doi: 10.2105/AJPH.2018.304673] [Medline: 30359112]
45. Han B, Compton WM, Blanco C, Crane E, Lee J, Jones CM. Prescription opioid use, misuse, and use disorders in U.S. adults: 2015 National Survey on Drug Use and Health. Ann Intern Med 2017 Sep 5;167(5):293-301. [doi: 10.7326/M17-0865] [Medline: 28761945]
46. Brooner RK, King VL, Kidorf M, Schmidt CW, Bigelow GE. Psychiatric and substance use comorbidity among treatment-seeking opioid abusers. Arch Gen Psychiatry 1997 Jan;54(1):71-80. [doi: 10.1001/archpsyc.1997.01830130077015] [Medline: 9006403]

47. Boyd AD, Gonzalez-Guarda R, Lawrence K, et al. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. J Am Med Inform Assoc 2023 Aug 18;30(9):1561-1566. [doi: 10.1093/jamia/ocad115] [Medline: 37364017]

Abbreviations

CDSS: clinical decision support system
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
EHR: electronic health record
***ICD*:** *International Classification of Diseases*
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
ML: machine learning
OUD: opioid use disorder

Edited by A Benis; submitted 20.10.23; peer-reviewed by L Zhang, X Dong; revised version received 15.03.24; accepted 20.04.24; published 04.06.24.

Please cite as:
Eguale T, Bastardot F, Song W, Motta-Calderon D, Elsobky Y, Rui A, Marceau M, Davis C, Ganesan S, Alsubai A, Matthews M, Volk LA, Bates DW, Rozenblum R
A Machine Learning Application to Classify Patients at Differing Levels of Risk of Opioid Use Disorder: Clinician-Based Validation Study
JMIR Med Inform 2024;12:e53625
URL: https://medinform.jmir.org/2024/1/e53625
doi:10.2196/53625

Impact of a Nationwide Medication History Sharing Program on the Care Process and End-User Experience in a Tertiary Teaching Hospital: Cohort Study and Cross-Sectional Study

Jungwon Cho[1,2], PhD; Sooyoung Yoo[3], PhD; Eunkyung Euni Lee[1,2,*], PharmD, PhD; Ho-Young Lee[3,4,*], MD, PhD

[1]

[2]

[3]

[4]

[*]these authors contributed equally

Corresponding Author:
Ho-Young Lee, MD, PhD

Abstract

Background: Timely and comprehensive collection of a patient's medication history in the emergency department (ED) is crucial for optimizing health care delivery. The implementation of a medication history sharing program, titled "Patient's In-home Medications at a Glance," in a tertiary teaching hospital aimed to efficiently collect and display nationwide medication histories for patients' initial hospital visits.

Objective: As an evaluation was necessary to provide a balanced picture of the program, we aimed to evaluate both care process outcomes and humanistic outcomes encompassing end-user experience of physicians and pharmacists.

Methods: We conducted a cohort study and a cross-sectional study to evaluate both outcomes. To evaluate the care process, we measured the time from the first ED assessment to urgent percutaneous coronary intervention (PCI) initiation from electronic health records. To assess end-user experience, we developed a 22-item questionnaire using a 5-point Likert scale, including 5 domains: information quality, system quality, service quality, user satisfaction, and intention to reuse. This questionnaire was validated and distributed to physicians and pharmacists. The Mann-Whiteny *U* test was used to analyze the PCI initiation time, and structural equation modeling was used to assess factors affecting end-user experience.

Results: The time from the first ED assessment to urgent PCI initiation at the ED was significantly decreased using the patient medication history program (mean rank 42.14 min vs 28.72 min; Mann-Whitney U=346; P=.03). A total of 112 physicians and pharmacists participated in the survey. Among the 5 domains, "intention to reuse" received the highest score (mean 4.77, SD 0.37), followed by "user satisfaction" (mean 4.56, SD 0.49), while "service quality" received the lowest score (mean 3.87, SD 0.79). "User satisfaction" was significantly associated with "information quality" and "intention to reuse."

Conclusions: Timely and complete retrieval using a medication history-sharing program led to an improved care process by expediting critical decision-making in the ED, thereby contributing to value-based health care delivery in a real-world setting. The experiences of end users, including physicians and pharmacists, indicated satisfaction with the program regarding information quality and their intention to reuse.

(JMIR Med Inform 2024;12:e53079) doi:10.2196/53079

KEYWORDS

health information system; HIS; medication history; history; histories; patients' own medication; satisfaction; DeLone and McLean Model of information systems success; value-based health care; emergency department; information system; information systems; emergency; urgent; drug; drugs; pharmacy; pharmacies; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutics; pharmaceuticals; pharmaceutical; medication; medications; sharing; user experience; survey; surveys; intention; intent; experience; experiences; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; acceptance; adoption

Introduction

Health information systems (HISs) play a vital role in the delivery of health care services, as they provide access to the patient's medical records, help track treatment progress, and support health care providers in making care decisions [1-3]. Although the development of HISs has revolutionized the provision of patient care and handling of patients' health information, in the transitional period toward the era of the fourth industrial revolution, studies that evaluate humanistic outcomes as well as clinical or economic outcomes caused by

HIS, are needed [4]. As leaders in health care settings have made various investments, such as time, money, and manpower, in managing HIS [5], the multifaceted evaluation of whether end users can use the HIS skillfully and achieve satisfaction in functionality and usability would be increasingly important in the future [4,6].

Health care organizations can ensure effective HIS use and improve the quality of patient care by conducting evaluations of HISs. These evaluations could allow health care organizations to proactively address issues related to system performance, integration, and data accuracy. However, evaluating the diversity and complexity of HISs in real-world clinical settings is a significant challenge [5,7]. Hospitals use different HISs depending on their work process, and the program related to direct patient care, including documentation and retrieval of medical records, or clinical decision support systems varies [8-10]. In addition, health care environments are constantly evolving with the emergence of innovative technologies [11]. Newly developed information systems or programs tend to be integrated into homegrown HISs after establishing a fully electronic medical record system. Thus, although HIS evaluations reporting economic, clinical, and humanistic outcomes could provide a balanced picture of the comprehensive impact of the health care interventions implemented, comprehensive evaluations of HISs are rarely conducted [12].

Acquisition of patients' complete medication use history could greatly enhance medication management and support physicians in making informed decisions. Accurate and efficient compilation of information can be more important when time-sensitive clinical decisions and subsequent interventions are made [13], especially in the emergency department (ED). However, previous studies have demonstrated that accurate and timely collection of patients' medication histories is challenging especially in the ED for various reasons, including patients with altered mental status due to confusion or intoxication, patients taking multiple outpatient prescriptions, and first-time patients to the hospital [14-16]. Since the treatment plan would change depending on the medication history, the prompt and complete evaluation of the medication history is vital. The process of collecting medication history was also described as a labor-intensive process, often requiring manual retrieval of information from outside the hospital [17,18]. Thus, a medication history sharing program called "Patient's In-home Medications at a Glance" was developed and successfully launched within a homegrown HIS known as BESTCare in Seoul National University Bundang Hospital (SNUBH) on January 11, 2021. The program enabled health professionals to access the patients' nationwide medication history swiftly and accurately from the Healthcare Insurance Review and Assessment Service database in South Korea with added features about the patient instructions and the identification guide for each medication. The rate of identification of patients' medication history within 24 hours was significantly improved at the ED after the implementation of the program [19]. However, comprehensive evaluations of querying patient medication history were necessary to provide a balanced picture of the medication history program, as an HIS intervention could have had an impact not only on the care process but also on humanistic outcomes, such as end-user experience about its functionality and usability, which may evolve over time.

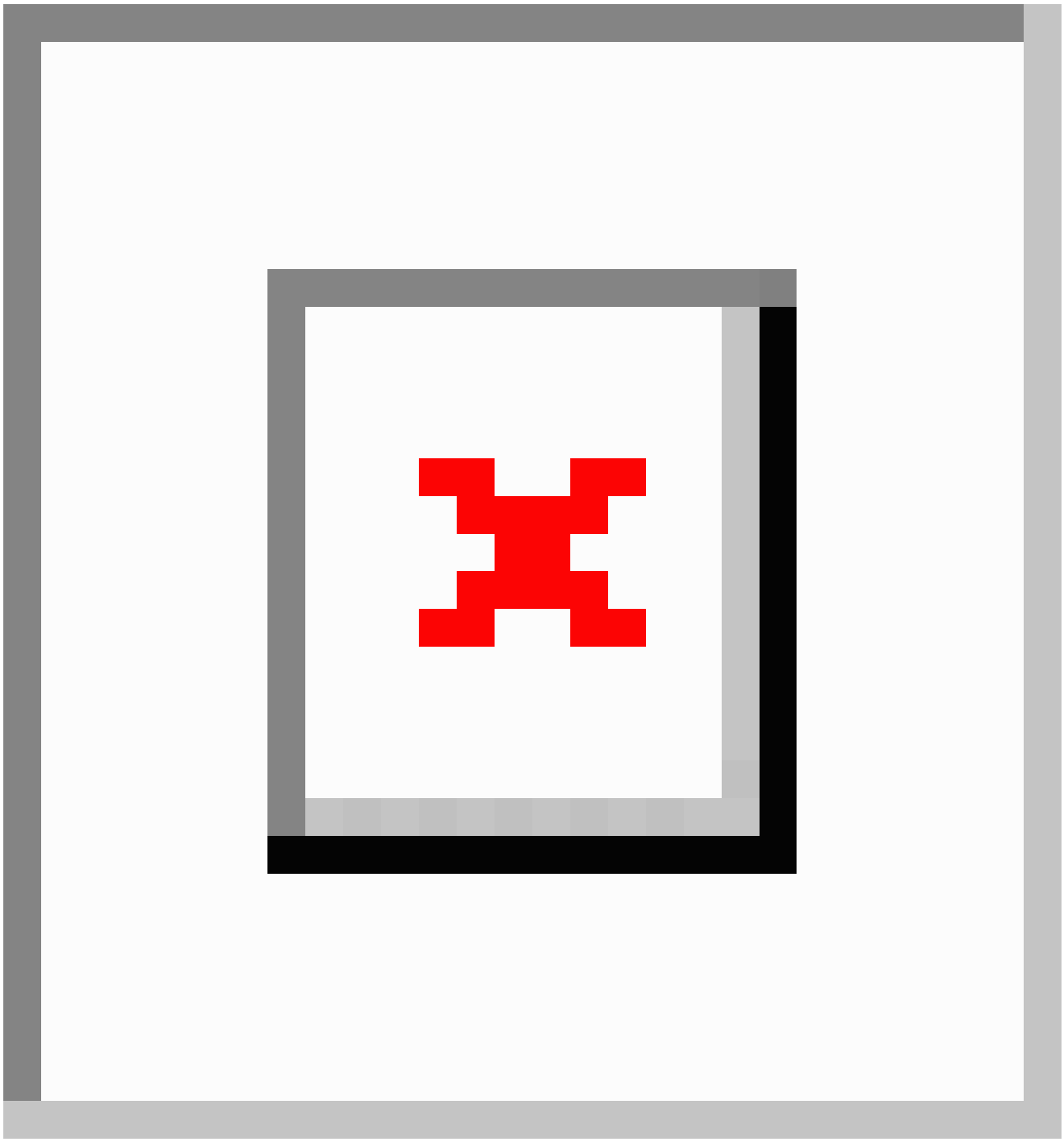
Therefore, this study aimed to evaluate the impact of an HIS intervention on health care delivery, namely medication history retrieval, using the "Patient's In-home Medications at a Glance" program. Specifically, we evaluated the care process outcome, that is, the time from the first ED assessment to urgent percutaneous coronary intervention (PCI) initiation, and the humanistic outcome, that is, the end-user experience among physicians and pharmacists.

Methods

Study Design

We conducted a cohort study and a cross-sectional study to evaluate both outcomes. We evaluated the impact of medication history retrieval using the "Patient's In-home Medications at a Glance" program on two aspects: (1) the care process outcome and (2) the end-user experience among physicians and pharmacists. Figure 1 shows the ED process and medication history check to describe the 2 outcomes of this study.

Figure 1. Emergency department (ED) process and medication history check depicting two outcomes: (1) time from the first ED assessment to urgent percutaneous coronary intervention (PCI) initiation as the care process outcome and (2) the end-user experience among physicians and pharmacists using the program as a humanistic outcome. Delayed medication history checks could increase the time of PCI initiation at the ED, especially in urgent clinical situations. The "Patient's In-home Medications at a Glance" program linking to the nationwide personal medication records provides more rapid and complete collections of medication history compared to manual retrievals that often require interviews with patients or caregivers at the ED (icons are made by Freepik).



First, we analyzed the care process to determine whether physicians' use of the program could expedite the time from the first ED assessment to urgent PCI initiation. Second, to assess end-user experience, we developed a questionnaire consisting of 22 survey items that were validated. We then conducted a website-based survey among physicians and pharmacists who served as end users of the program.

Care Process Outcome

Data Collection

For the care process, patients who were admitted to the ED for the first time from January 1, 2021, to December 31, 2022, were included to estimate the impact of the program on the collection of patients' drug therapy. The outcome was defined as the time of initiating urgent PCI after the first assessment by ED physicians from January 1, 2021, to December 31, 2022. Urgent PCI was defined as PCI performed within an hour of admission

to the ED. As the identification of the patient's medication use history was required to further improve the care plan, the time from the first ED assessment to urgent PCI initiation was analyzed.

Data Analysis

To analyze the impact of the program on the care process, data were extracted from the SNUBH electronic database. We performed a Mann-Whitney *U* test to evaluate the difference in the time from the first ED assessment to urgent PCI initiation between patients who were queried about their medication use history by physicians via the program and those who were not.

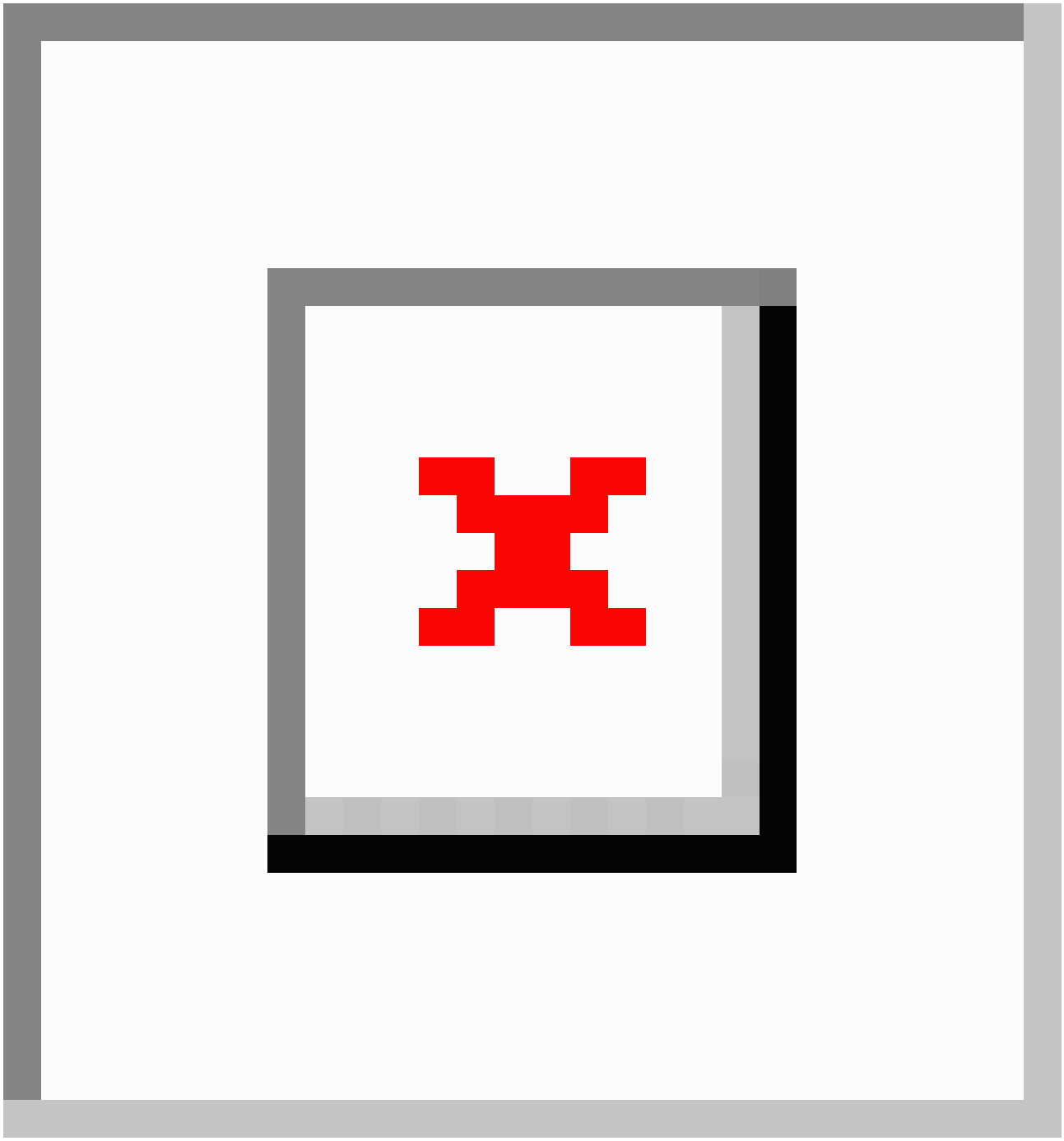
All analyses were performed using IBM SPSS Statistics (version 22.0; IBM Corp) and R (version 4.0.2; R Foundation for Statistical Computing).

Survey and Assessing Factors Affecting End-User Experience on the Program

Survey Development With a Conceptual Framework

To assess end-user experience and whether end users are satisfied with HIS and their intention to reuse it, we adopted the updated DeLone and McLean Model of Information Systems Success (DMISM) [20] for survey development. The updated DMISM provides a conceptual framework to suggest the factors necessary for the provision of use and benefits from the HIS. Based on the updated DMISM, we proposed that the quality of the information system consists of 3 quality domains: information quality, system quality, and service quality. These domains are necessary for user satisfaction and are instrumental in driving users' intentions to reuse the system. In this study, we narrowed the scope to physicians and pharmacists who were already using the program. Therefore, we adjusted the factor of "intention to use" and "use" in the updated DMISM to "intention to reuse." Due to the nature of the HIS, "intention to reuse" of the program by end users is considered the ultimate and crucial goal. By setting it as the final outcome variable, "intention to reuse" is influenced by preceding user satisfaction. Therefore, we established the research model with the relationship that "user satisfaction" affects "intention to reuse." These domains were used to develop the survey (Figure 2).

Figure 2. A research model for the survey development. The updated DeLone and McLean Model of Information Systems Success [20] provides domain variables consisting of 3 quality domains (ie, information, system, and service quality) as well as outcome domains.



We collected 32 survey questionnaires that assessed each quality domain regarding previous studies [3,21-24]. Through face validation with 6 pharmacists, a physician, and a medical informatics professor every 2 weeks for 3 months, the survey questionnaires were classified according to each domain. The questionnaires were eliminated or revised to reflect the contextual significance of the program. The draft survey finally consisted of 22 questionnaires, and a pilot study was conducted with 10 pharmacists and 2 physicians at SNUBH.

The survey was conducted from December 15, 2022, to December 28, 2022, at SNUBH. We used a web-based survey to collect data on the end-user experience efficiently and rapidly. The survey link was distributed to all physicians and pharmacists at the hospital via email. Survey completion was expected to take approximately 5 minutes. The items in the survey were rated on a 5-point Likert scale (1=not at all; 5=very much). Only those who provided consent after receiving an explanation of the background and purpose of the survey were included.

Data Analysis

An exploratory factor analysis of the results was then performed to determine how the items were classified into components. We used the Kaiser-Meyer-Olkin measure to assess sampling adequacy and obtained a specific value of 0.858, surpassing the recommended threshold of 0.5. The suitability of the data for factor analysis was further confirmed through the Bartlett test

of sphericity, yielding a statistically significant result (χ^2_{105}=723.6; *P*<.001). The analysis of communality, indicating the explanatory power between measurement variables and extracted factors, was performed. Considering the general criterion that variables with communality below 0.4 are deemed low and should be excluded from factor analysis, 8 questions were excluded. Consequently, 14 questionnaires were retained (Table S1 in Multimedia Appendix 1).

Subsequently, we conducted a reliability analysis of the survey items and calculated Cronbach α. We analyzed the convergent and discriminant validity of the constructs. We used SPSS to conduct statistical analyses, including factor and reliability analyses. Finally, structural equation modeling (SEM) was used to evaluate the structural correlations among the domains using the AMOS 25 software (version 25.0; IBM Corp). SEM was chosen to provide a comprehensive understanding of the relationships among survey variables and to help validate the theoretical models with a visual representation.

Ethical Considerations

This study was approved by the Institutional Review Board of SNUBH (B-2203-746-001; April 21, 2022), and the requirement of obtaining written consent was waived, as this study did not contain sensitive personally identifiable information.

Results

Care Process Outcome

Of the 162 patients who were admitted to the ED and visited the hospital for the first time over a 2-year period, 77 who underwent urgent PCIs within an hour from the first ED assessment to urgent PCI initiation were included. Patients who were regularly visiting hospitals with chronic diseases were excluded. Table 1 describes the demographic characteristics of patients, including gender, age, department, tests, and diagnosis, between the patient group (n=59), for which the doctor did not use the program, and the patient group (n=18), whose medications were accessed through the program.

Table . Demographics of patients receiving urgent percutaneous coronary intervention by use of the medication history program during the study period at an emergency department (n=77[a]).

Characteristics		No (without the program; n=59)	Yes (with the program; n=18)
Sex (male), n (%)		50 (84.7)	11 (61.1)
Age (years), mean (SD)		64.3 (12.1)	68.9 (12.4)
Department at discharge, n (%)			
	Cardiology	54 (91.5)	16 (88.9)
	Others	5 (8.5)	2 (11.1)
Had CT[b] scan, n (%)		12 (20.3)	3 (16.7)
Diagnosis, n (%)			
	ST elevation myocardial infarction	53 (89.8)	16 (88.9)
	Others	10 (16.9)	4 (22.2)

[a]Patients receiving percutaneous coronary intervention within an hour at an emergency department from January 12, 2021, to December 31, 2022.

[b]CT: computed tomography.

Changes in time from the first ED assessment to urgent PCI initiation significantly decreased in patients who used the program (n=18; mean rank 28.72 min) versus patients who did not use the program (n=59; mean rank 42.14 min; Mann-Whitney *U*=346; *P*=.03).

Survey and Assessing Factors Affecting End-User Experience on the Program

Survey Participants' Characteristics

During the 2-week survey period, we received survey responses from 112 participants in the hospital. Among them, we removed the responses of 10 participants who never used the "Patient's In-home Medication at a Glance" based on their answers to the first question. In addition, the responses of 5 participants who gave the same rating to the negative and positive questions were removed, as they were considered either not meaningful or not sincere to the survey, leaving 97 responses for analysis. Table 2 presents the characteristics. Participants included 62 (63.9%) physicians and 35 (36.1%) pharmacists, and the mean use count during the week was approximately 10.8 (SD 13.9).

Table . Participants' characteristics (N=97).

Characteristics				Values, n (%)
Occupation				
	Physician (n=62, 63.9%)			
		Position		
			Professor	35 (56.5)
			Resident	27 (43.5)
		Department		
			Internal medicine	50 (80.6)
			Surgery	12 (19.3)
		Workplace		
			Ambulatory clinic	24 (38.7)
			General ward	21 (33.9)
			Emergency room	11 (17.7)
			Intensive care unit	6 (9.7)
	Pharmacist			35 (36.1)
EHR [a] experience (years)				
	1			5 (5.2)
	3			20 (20.6)
	5			22 (22.7)
	10			20 (20.6)
	>10			30 (30.9)
Sex				
	Male			32 (33.0)
	Female			65 (67.0)
Age (years)				
	≤30			14 (14.4)
	31-40			58 (59.8)
	41-50			20 (20.6)
	>50			5 (5.2)
Weekly frequency of using the program				
	Mean (SD)			10.7 (13.9)
	Median (IQR)			6 (4-10)

[a]EHR: electronic health record.

Evaluation of the Survey Results

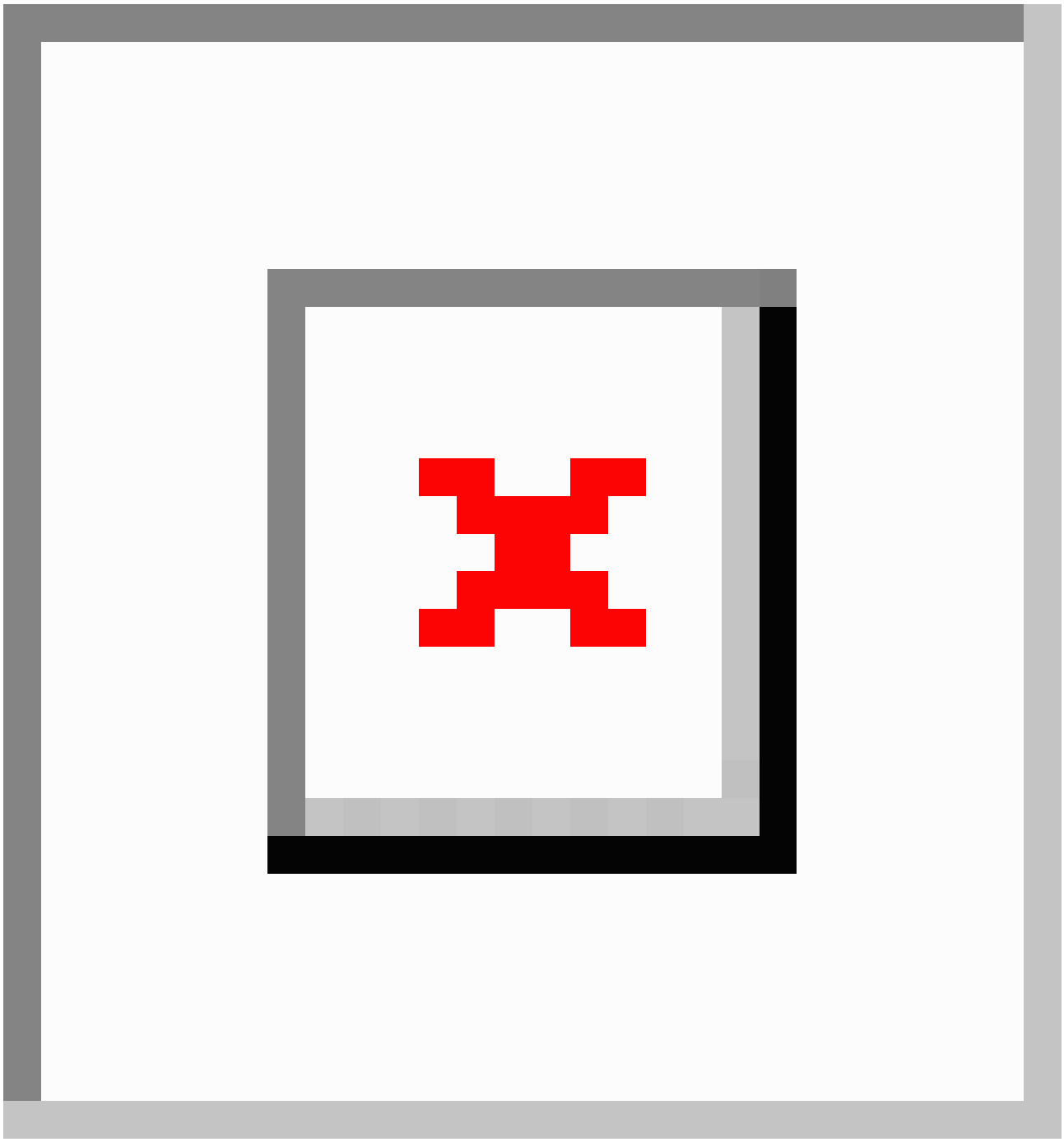
Of the 22 survey questions, the updated DMISM comprised 14 questions in 5 domains. After performing exploratory factor analysis, we calculated the mean score of each domain and Cronbach α to confirm the consistency of the items. This reliability analysis revealed that Cronbach α for all variables exceeded 0.80 (information quality: 0.808; system quality: 0.834; and service quality: 0.800), except for user satisfaction (Cronbach α=0.788) and intention to use (Cronbach α=0.795).

On a 5-point scale, the mean scores values for the information, system, and service quality of the program were 4.11 (SD 0.76), 4.24 (SD 0.75), and 3.87 (SD 0.79), respectively. User satisfaction (4.56, SD 0.49) and intention to reuse (4.77, SD 0.37) were measured. Among the 5 domains of the survey questionnaire, intention to reuse obtained the highest score. The estimates and weights of all 5 domains were analyzed, and no issues were observed in the convergent validity of the constructs (Table S2 in Multimedia Appendix 1). In addition, the subsequent analysis revealed the absence of discriminant validity (Table S3 in Multimedia Appendix 1).

Structural Correlations Between Domains

The SEM images are shown in Figure 3. The model fit indices were calculated as follows: χ^2_{70}=103.413 (P<.001); goodness-of-fit index=0.868 (recommended: 0-1.0); root mean square residual=0.039 (recommended: 0-0.05); and root mean square error of approximation=0.071 (recommended: 0.05-0.08). The comparative fit index and the Tucker-Lewis index for the model exceeded 0.9. The normed fit index and adjusted goodness-of-fit index values were lower than the recommended values of 0.859 and 0.802, respectively. Thus, this model was confirmed to be appropriate for assessing the factors affecting the “intention to reuse” program as an end-user experience.

Figure 3. Results of the research model using the structural equation modeling analysis. Significant paths are indicated with solid lines, while nonsignificant paths are shown with dotted lines. Two significant paths are shown: information quality toward user satisfaction and user satisfaction toward intention to reuse. The standardized beta values are presented. *P=.01; ***P<.001. CFI: Comparative Fit Index; RMSEA: root mean square error of approximation; TLI: Tucker-Lewis Index.



The associations between the latent variables were positive, supporting our hypotheses. Among the 3 quality domains, “information quality” had a significantly positive influence on “user satisfaction.” Consequently, the influence of “information quality” in “user satisfaction” and the influence of “user satisfaction” in “intent to reuse” were significantly associated.

Discussion

Principal Results

This study aimed to evaluate the impact of medication history retrieval using the "Patient's In-home Medications at a Glance" program in homegrown HISs during the 2-year maintenance phase after program implementation. The significance of our findings was twofold. First, we conducted a comprehensive evaluation of the impact of the nationwide medication history-sharing program, consisting of care process outcomes and end-user experiences as humanistic outcomes. We elaborately planned both the care process and humanistic outcomes of 2-year use, which allowed the program to stabilize, after its implementation in the HIS [23]. The care process, focusing on the time required for urgent PCI initiation, was improved in the patient group, whose physicians used the program and experienced expedited urgent PCI initiation. Thus, the use of the program could help identify whether patients are taking an antiplatelet or anticoagulant agent when they are unconscious or are unable to identify their medications. Regarding humanistic outcomes, the survey showed high scores overall, especially for "user satisfaction" and "intention to reuse." The increasing trend in the use of the "Patient's In-home Medications at a Glance" program by physicians and pharmacists indicates the successful integration of the newly developed program into the HIS, as evidenced by a positive end-user experience.

Second, we assessed factors affecting end-user experience using SEM; "information quality" significantly influenced "user satisfaction," and "user satisfaction," in turn, positively enhanced "intention to reuse." Since the survey was developed with the updated DMISM, which is a conceptual framework to suggest factors necessary for the "intention to reuse" the program, we could examine whether and how the 3 quality domains, including information, system, and service, affect "user satisfaction" and how "user satisfaction" affects "intention to reuse." These findings highlight the potential of the HIS in supporting clinical decision-making and contributing to value-based health care through the provision of a comprehensive medication use history.

Implications

Value-based health care is an approach to health care delivery in which providers are paid based on the patient's health outcomes [25], while reducing costs [26]. The benefits of a value-based health care system include reduced treatment costs, increased care efficiency, and reduced risks [27]. Measuring a patient's clinical outcomes is a major aim of value-based health care. In our study, we measured both care process outcomes and end-user experiences, which help present humanistic outcomes. Hence, a comprehensive evaluation was conducted by selecting both outcomes to determine the impact of the interventions using the HIS. Health service providers should provide patient-centered team care, share patients' medical information, and measure the care process using the HIS. The physicians were able to collect the patients' complete medication use histories in a friendly manner, even if the patients were unable to identify the exact medications they were taking. As access to a complete medication use history could help physicians make clinical decisions and collaborate care within the hospital [28], the HIS could help improve the patient's outcomes. Thus, HISs can play a vital role in value-based health care by delivering comprehensive and up-to-date information, including medication use history, laboratory results, and other medical records.

In terms of the association between the survey domains, the updated DMISM was applied to identify the quality factors that contribute to "user satisfaction," which affects end users' "intention to reuse." According to Alzahrani et al [29], 3 quality domains are significantly related to "user satisfaction" and "intention to reuse" and consequently affect actual usage. By conducting an SEM analysis of the survey results, our model revealed a significant effect of "information quality" on "user satisfaction," as well as "user satisfaction" on "intention to reuse." These results indicate that providing complete, accurate, and regent information is important for "user satisfaction," ultimately driving the "intention to reuse." A previous study stated that studies assessing the acceptance of HISs have been conducted from the physicians' perspective, not the clinical pharmacists' [30]. Since the program has been used by physicians and pharmacists, we could assess the factors affecting end-user experience in both professional groups. If the quality of information in an HIS is not guaranteed, health care professionals will not use specific programs in the HIS.

Limitations

This study had some limitations. First, we developed and implemented the "Patient's In-home Medications at a Glance" program in a single hospital. Thus, outcomes, such as care processes or factors affecting end-user experience, cannot be generalized to other hospitals in South Korea. However, as the Healthcare Insurance Review and Assessment Service has established guidelines for program development, further studies that use similar HISs could be conducted in other hospitals. Second, the pretest and posttest studies had the inherent limitations of nonrandomized, uncontrolled study designs. Although we showed the impact of the program on the time to PCI as the care process, we could not capture the long-term effects on clinical outcomes, such as survival rates or extended hospital stays. Nevertheless, our findings regarding the care process, specifically the reduction in time from the first ED assessment to urgent PCI initiation, could be meaningful not only in expediting clinical decisions but also in the evaluation of HISs in a real-world health care setting. Third, a notable limitation of our study is the imbalanced distribution of participants between the patient groups with or without the program (18 vs 59 participants) and the small number of patients in the group using the program. This uneven and small sample size raises concerns about the statistical robustness of our findings. Future research endeavors should prioritize achieving a more equitable number and distribution of patients to enhance the reliability and generalizability of our conclusions. Although our study offers valuable insights, the limitation of uneven and small sample sizes underscores the importance of cautious interpretation and highlights a potential area for improvement in subsequent research. Fourth, in the results of the SEM analysis, "information quality" was a standalone significant

factor among 3 quality domains influencing "user satisfaction." It is possible that the developed survey item may not adequately address the measurement of the quality domain. Lastly, our focus in this study was on system acceptability rather than the direct improvement in the health of the patients. We plan to focus more on the clinical outcome of the program, which includes not only medication information but also ensuring comprehensive disease management. This approach should be followed up for future measurements in subsequent studies.

Conclusions

Our findings highlight the impact of the rapid and complete medication history retrieval using the "Patient's In-home Medications at a Glance" program on the care process and end-user experience. A significantly positive effect was found on the care process by expediting urgent PCI initiation time at the ED, thereby contributing to value-based healthcare delivery in a real-world setting. Moreover, the HIS intervention provided high-quality information to physicians and pharmacists, resulting in high satisfaction. Long-term assessments can provide valuable insights into the sustained impact of the program, further optimizing patient outcomes.

Acknowledgments

We would like to thank the project team and program developer of the Medical Informatics Team for developing and implementing the program and assisting with data retrieval. We are grateful to all the pharmacists and physicians who participated in patient care. We would like to acknowledge the contributions of EEL as a cocorresponding author. We would also like to thank the Brain Korea (BK) 21 Plus Project of the National Research Foundation of Korea.

Authors' Contributions

JC, SY, HYL, and EEL contributed to the conception and design of the research, the acquisition and analysis of the data, as well as the interpretation of the data. They also drafted the manuscript. All authors critically revised the manuscript, agreed to be fully accountable for ensuring the integrity and accuracy of the work, and read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey items, convergent validity, and discriminant validity.
[DOCX File, 20 KB - medinform_v12i1e53079_app1.docx]

References

1. Ludwick DA, Doucette J. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. Int J Med Inform 2009 Jan;78(1):22-31. [doi: 10.1016/j.ijmedinf.2008.06.005] [Medline: 18644745]
2. Aggelidis VP, Chatzoglou PD. Hospital information systems: measuring end user computing satisfaction (EUCS). J Biomed Inform 2012 Jun;45(3):566-579. [doi: 10.1016/j.jbi.2012.02.009] [Medline: 22426283]
3. Ojo AI. Validation of the DeLone and McLean information systems success model. Healthc Inform Res 2017 Jan;23(1):60-66. [doi: 10.4258/hir.2017.23.1.60] [Medline: 28261532]
4. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: an evaluation of information system impact and contingency factors. Int J Med Inform 2014 Nov;83(11):779-796. [doi: 10.1016/j.ijmedinf.2014.06.011] [Medline: 25085286]
5. Laukka E, Huhtakangas M, Heponiemi T, Kanste O. Identifying the roles of healthcare leaders in HIT implementation: a scoping review of the quantitative and qualitative evidence. Int J Environ Res Public Health 2020 Apr 21;17(8):2865. [doi: 10.3390/ijerph17082865] [Medline: 32326300]
6. Jones SS, Rudin RS, Perry T, Shekelle PG. Health information technology: an updated systematic review with a focus on meaningful use. Ann Intern Med 2014 Jan 7;160(1):48-54. [doi: 10.7326/M13-1531] [Medline: 24573664]
7. Joseph AL, Stringer E, Borycki EM, Kushniruk AW. Evaluative frameworks and models for health information systems (HIS) and health information technologies (HIT). Stud Health Technol Inform 2022 Jan 14;289:280-285. [doi: 10.3233/SHTI210914] [Medline: 35062147]
8. Bates DW, Gawande AA. Improving safety with information technology. N Engl J Med 2003 Jun 19;348(25):2526-2534. [doi: 10.1056/NEJMsa020847] [Medline: 12815139]
9. Mogharbel A, Dowding D, Ainsworth J. Physicians' use of the computerized physician order entry system for medication prescribing: systematic review. JMIR Med Inform 2021 Mar 4;9(3):e22923. [doi: 10.2196/22923] [Medline: 33661126]

10. Neame MT, Sefton G, Roberts M, Harkness D, Sinha IP, Hawcutt DB. Evaluating health information technologies: a systematic review of framework recommendations. Int J Med Inform 2020 Oct;142:104247. [doi: 10.1016/j.ijmedinf.2020.104247] [Medline: 32871491]
11. Zeadally S, Siddiqui F, Baig Z, Ibrahim A. Smart healthcare: challenges and potential solutions using Internet of Things (IOT) and big data analytics. PSU Res Rev 2019 Feb;4:149-168. [doi: 10.1108/PRR-08-2019-0027]
12. Gunter MJ. The role of the ECHO model in outcomes research and clinical practice improvement. Am J Manag Care 1999 Apr;5(4 Suppl):S217-S224. [Medline: 10387542]
13. Marshall J, Hayes BD, Koehl J, et al. Effects of a pharmacy-driven medication history program on patient outcomes. Am J Health Syst Pharm 2022 Sep 22;79(19):1652-1662. [doi: 10.1093/ajhp/zxac143] [Medline: 35596269]
14. Cadwallader J, Spry K, Morea J, Russ AL, Duke J, Weiner M. Design of a medication reconciliation application: facilitating clinician-focused decision making with data from multiple sources. Appl Clin Inform 2013 Mar 13;4(1):110-125. [doi: 10.4338/ACI-2012-12-RA-0057] [Medline: 23650492]
15. Cornish PL, Knowles SR, Marchesano R, et al. Unintended medication discrepancies at the time of hospital admission. Arch Intern Med 2005 Feb 28;165(4):424-429. [doi: 10.1001/archinte.165.4.424] [Medline: 15738372]
16. Hart C, Price C, Graziose G, Grey J. A program using pharmacy technicians to collect medication histories in the emergency department. P T 2015 Jan;40(1):56-61. [Medline: 25628508]
17. Lau HS, Florax C, Porsius AJ, De Boer A. The completeness of medication histories in hospital medical records of patients admitted to general internal medicine wards. Br J Clin Pharmacol 2000 Jun;49(6):597-603. [doi: 10.1046/j.1365-2125.2000.00204.x] [Medline: 10848724]
18. Tamblyn R, Huang AR, Meguerditchian AN, et al. Using novel Canadian resources to improve medication reconciliation at discharge: study protocol for a randomized controlled trial. Trials 2012 Aug 27;13:150. [doi: 10.1186/1745-6215-13-150] [Medline: 22920446]
19. Cho J, Lee E, Lee K, Lee HY, Lee E. Continuity of care with a one-click medication history program: patient's in-home medications at a glance. Int J Med Inform 2022 Apr;160:104710. [doi: 10.1016/j.ijmedinf.2022.104710] [Medline: 35183048]
20. William HD, Ephraim RM. The DeLone and McLean model of information systems success: a ten-year update. J Manag Info Syst 2003 Apr;19(4):9-30. [doi: 10.1080/07421222.2003.11045748]
21. Cho HH. Study on influence of perceived quality factor of smartphone on satisfaction & continued use intention - from the standpoint of updated DeLone & McLean's information system success model -. Entrue J Inf Technol 2012 Aug;11(2):167-180 [FREE Full text]
22. Shim M, Jo HS. What quality factors matter in enhancing the perceived benefits of online health information sites? application of the updated DeLone and McLean information systems success model. Int J Med Inform 2020 May;137:104093. [doi: 10.1016/j.ijmedinf.2020.104093] [Medline: 32078918]
23. Bossen C, Jensen LG, Udsen FW. Evaluation of a comprehensive EHR based on the DeLone and McLean model for IS success: approach, results, and success factors. Int J Med Inform 2013 Oct;82(10):940-953. [doi: 10.1016/j.ijmedinf.2013.05.010] [Medline: 23827768]
24. Song T, Deng N, Cui T, et al. Measuring success of patients' continuous use of mobile health services for self-management of chronic conditions: model development and validation. J Med Internet Res 2021 Jul 13;23(7):e26670. [doi: 10.2196/26670] [Medline: 34255685]
25. NEJM Catalyst. What is value-based healthcare? NEJM Catalyst. 2017 Jan 1. URL: https://catalyst.nejm.org/doi/full/10.1056/CAT.17.0558 [accessed 2023-03-01]
26. Ibanez-Sanchez G, Fernandez-Llatas C, Martinez-Millana A, et al. Toward value-based healthcare through interactive process mining in emergency rooms: the stroke case. Int J Environ Res Public Health 2019 May 20;16(10):1783. [doi: 10.3390/ijerph16101783] [Medline: 31137557]
27. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769. [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
28. Vargas V, Blakeslee WW, Banas CA, Teter C, Dupuis-Dobson K, Aboud C. Use of complete medication history to identify and correct transitions-of-care medication errors at psychiatric hospital admission. PLoS One 2023;18(1):e0279903. [doi: 10.1371/journal.pone.0279903] [Medline: 36696376]
29. Alzahrani AI, Mahmud I, Ramayah T, Alfarraj O, Alalwan N. Modelling digital library success using the DeLone and McLean information system success model. J Librariansh Inf Sci 2019 Jun;51(2):291-306. [doi: 10.1177/0961000617726123]
30. English D, Ankem K, English K. Acceptance of clinical decision support surveillance technology in the clinical pharmacy. Inform Health Soc Care 2017 Mar;42(2):135-152. [doi: 10.3109/17538157.2015.1113415] [Medline: 26890621]

Abbreviations

DMISM: DeLone and McLean Model of Information Systems Success
ED: emergency department
HIS: health information system
PCI: percutaneous coronary intervention

SEM: structural equation modeling
SNUBH: Seoul National University Bundang Hospital

Edited by C Lovis; submitted 25.09.23; peer-reviewed by D Carvalho, G Vergeire-Dalmacion; revised version received 16.01.24; accepted 04.02.24; published 20.03.24.

Please cite as:
Cho J, Yoo S, Lee EE, Lee HY
Impact of a Nationwide Medication History Sharing Program on the Care Process and End-User Experience in a Tertiary Teaching Hospital: Cohort Study and Cross-Sectional Study
JMIR Med Inform 2024;12:e53079
URL: https://medinform.jmir.org/2024/1/e53079
doi:10.2196/53079

A Patient Similarity Network (CHDmap) to Predict Outcomes After Congenital Heart Surgery: Development and Validation Study

Haomin Li[1,*], PhD; Mengying Zhou[1,2,*], MSc; Yuhan Sun[1,2,*], BSc; Jian Yang[1,2], BSc; Xian Zeng[1,2], PhD; Yunxiang Qiu[3], MD; Yuanyuan Xia[3], MD; Zhijie Zheng[3], MD; Jin Yu[4], MD; Yuqing Feng[1], MSc; Zhuo Shi[5], MD; Ting Huang[5], MD; Linhua Tan[3], MD; Ru Lin[5], MD; Jianhua Li[5], MD; Xiangming Fan[5], MD; Jingjing Ye[4], MD; Huilong Duan[2], PhD; Shanshan Shi[3,*], MD; Qiang Shu[5,*], MD

1

2

3

4

5

[*]these authors contributed equally

Corresponding Author:
Haomin Li, PhD

Abstract

Background: Although evidence-based medicine proposes personalized care that considers the best evidence, it still fails to address personal treatment in many real clinical scenarios where the complexity of the situation makes none of the available evidence applicable. “Medicine-based evidence” (MBE), in which big data and machine learning techniques are embraced to derive treatment responses from appropriately matched patients in real-world clinical practice, was proposed. However, many challenges remain in translating this conceptual framework into practice.

Objective: This study aimed to technically translate the MBE conceptual framework into practice and evaluate its performance in providing general decision support services for outcomes after congenital heart disease (CHD) surgery.

Methods: Data from 4774 CHD surgeries were collected. A total of 66 indicators and all diagnoses were extracted from each echocardiographic report using natural language processing technology. Combined with some basic clinical and surgical information, the distances between each patient were measured by a series of calculation formulas. Inspired by structure-mapping theory, the fusion of distances between different dimensions can be modulated by clinical experts. In addition to supporting direct analogical reasoning, a machine learning model can be constructed based on similar patients to provide personalized prediction. A user-operable patient similarity network (PSN) of CHD called CHDmap was proposed and developed to provide general decision support services based on the MBE approach.

Results: Using 256 CHD cases, CHDmap was evaluated on 2 different types of postoperative prognostic prediction tasks: a binary classification task to predict postoperative complications and a multiple classification task to predict mechanical ventilation duration. A simple poll of the *k*-most similar patients provided by the PSN can achieve better prediction results than the average performance of 3 clinicians. Constructing logistic regression models for prediction using similar patients obtained from the PSN can further improve the performance of the 2 tasks (best area under the receiver operating characteristic curve=0.810 and 0.926, respectively). With the support of CHDmap, clinicians substantially improved their predictive capabilities.

Conclusions: Without individual optimization, CHDmap demonstrates competitive performance compared to clinical experts. In addition, CHDmap has the advantage of enabling clinicians to use their superior cognitive abilities in conjunction with it to make decisions that are sometimes even superior to those made using artificial intelligence models. The MBE approach can be embraced in clinical practice, and its full potential can be realized.

(JMIR Med Inform 2024;12:e49138) doi:10.2196/49138

KEYWORDS

medicine-based evidence; general prediction model; patient similarity; congenital heart disease; echocardiography; postoperative complication; similarity network; heart; cardiology; NLP; natural language processing; predict; predictive; prediction; complications; complication; surgery; surgical; postoperative

Introduction

Congenital heart disease (CHD) is the most common type of birth defect, with birth prevalence reported to be 1% of live births worldwide [1]. Despite remarkable success in the surgical and medical management that has increased the survival of children with CHD [2], the quality of treatment and prognosis after congenital heart surgery remains unsatisfactory and varies across centers [3,4]. The reason for this is that the complexity of the disease, clinical heterogeneity within lesions, and small number of patients with specific forms of CHD severely degrade the precision and value of estimates of average treatment effects provided by randomized controlled trials on the average patient. Some visionary researchers have proposed a new paradigm called "medicine-based evidence" (MBE), in which big data and machine learning techniques are embraced to interrogate treatment responses among appropriately matched patients in real-world clinical practice [5,6].

Postoperative complications in congenital heart surgery have been inconsistently reported but have important contributions to mortality, hospital stay, cost, and quality of life [7-9]. Heart centers with the best outcomes might not report fewer complications but rather have systems in place to recognize and correct complications before deleterious outcomes ensue [8]. The early detection of deterioration after congenital heart surgery enables prompt initiation of therapy, which may result in reduced impairment and earlier rehabilitation. Several risk scoring systems, such as the Risk Adjustment for Congenital Heart Surgery 1 (RACHS-1) method, Aristotle score, and Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery (STS-EACTS) score, have been developed and used to adjust the risk of in-hospital morbidity and mortality [10-13]. However, most of these consensus-based risk models only focus on the procedures themselves and ignore the differences between centers and patients. Specific patient characteristics, such as lower weight [14] and longer cardiopulmonary bypass time [15], especially the quantitative echocardiographic indicators used by clinicians to understand CHD conditions, were not incorporated into these models nor can they be adjusted for. Based on the increasing number of CHD databases being built, some machine learning–based predictive models have recently been used to identify independent risk factors and predict complications after congenital heart surgery [16-18]. These predictive models achieved outstanding performance compared to traditional risk scores, but these models are usually only capable of performing a single task. In addition, such models often contain hundreds of features, so for clinicians, understanding how to interpret the prediction from a complicated machine learning model is still a challenge [19]. Based on our previous studies [16-18], as the model becomes more complex and more variables are included, the results are better, but it is more difficult to understand and accept clinically. Although some explainable artificial intelligence (AI) techniques continue to evolve [20,21], machine learning prediction models are still a black box for clinicians. Due to the lack of understanding and manipulation of the model, clinicians often lack confidence in the predicted outcomes, which severely hampers the entry of these machine learning models into routine care.

Patient similarity networks (PSNs) are an emerging paradigm for precision medicine, in which patients are clustered or classified based on their similarities in various features [22,23]. PSNs address many challenges in data analytics and is naturally interpretable. In a PSN, each node is an individual patient, and the distance (or edge) between 2 nodes corresponds to pairwise patient similarity for given features. PSNs naturally handle heterogeneous data, as any data type can be converted into a similarity network by defining similarity measures [24,25]. A PSN generated based on a large cohort of patients will show several subgroups of patients who are tightly connected. If a new patient is located on the PSN, neighbors that have similar features with known risk or prognosis will inform clinicians of the potential risk and prognosis of the patient. This mimics the clinical reasoning of many experienced clinical experts, who often relate a patient to similar patients they have seen. Moreover, representing patients by similarity is conceptually intuitive and explainable because it can convert the data into network views, where the decision boundary can be visually evident [26]. PSNs can also provide a feasible engineering solution for the MBE framework, which, based on a library of "approximate matches" consisting of a group of patients who share the greatest similarity with the index case, can be examined to estimate the effects of various treatments within the context of the individual patient's specific characteristics [6].

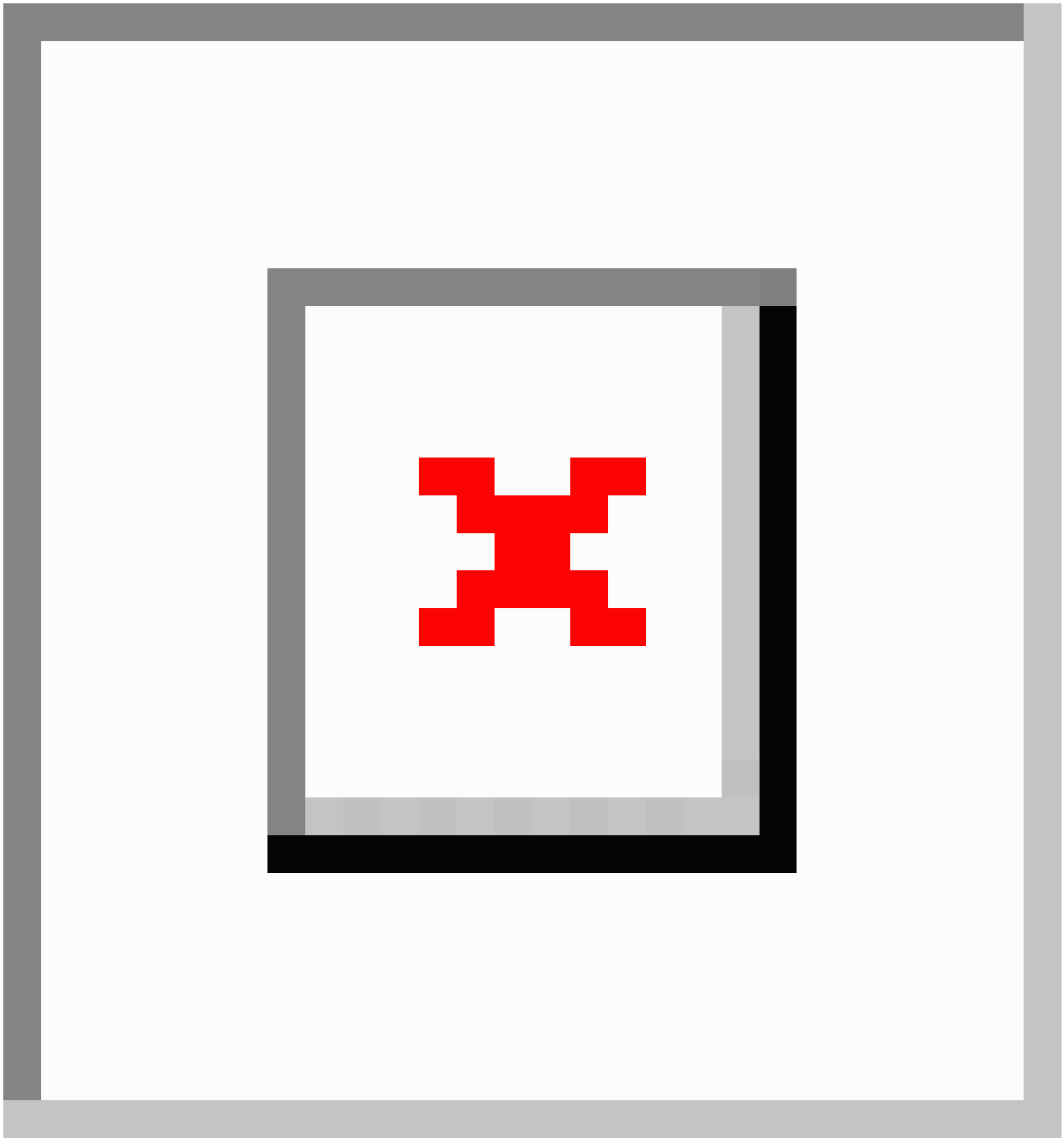
PSNs have been reported in many studies. Although early PSN studies have focused on using omics data in precision medicine [27-29], with the development of electronic health record (EHR) systems, abundant, complex, high-dimensional, and heterogeneous data are being captured during daily care, and some EHR-based patient similarity frameworks have been proposed for diagnosis [30], subgroup patients [31,32], outcome prediction [33], drug recommendation [34,35], and disease screening [36]. However, studies of PSNs that predict the outcome after CHD surgery have not been reported. A perspective article proposed an MBE conceptual framework for CHD [6], in which similarity analysis is used to generate a library of "approximate matches." However, they did not provide any technical solution for this framework. The challenge in applying PSNs in a real clinical setting is, first of all, to assess the distance between patients with complex conditions such as CHD in a computable way. However, mimicking clinical analogy reasoning is not a simple math formula based on various patients' attributes. The structure-mapping theory in cognitive science argues that advanced cognitive functions are involved in the analysis of relationship similarity above attribute similarity [37]. Analogy inference requires advanced cognitive activity, which current AI technology lacks but clinical experts are good at. However, all established models ignore this important feature of patient similarity analysis, in that it should not only measure patients' distance but also put clinicians back behind the wheel to generate MBE for clinical decision-making. In this study, we aimed to develop and evaluate a clinician-operable PSN of CHD to try to mitigate the above problems.

Methods

Study Design and Population

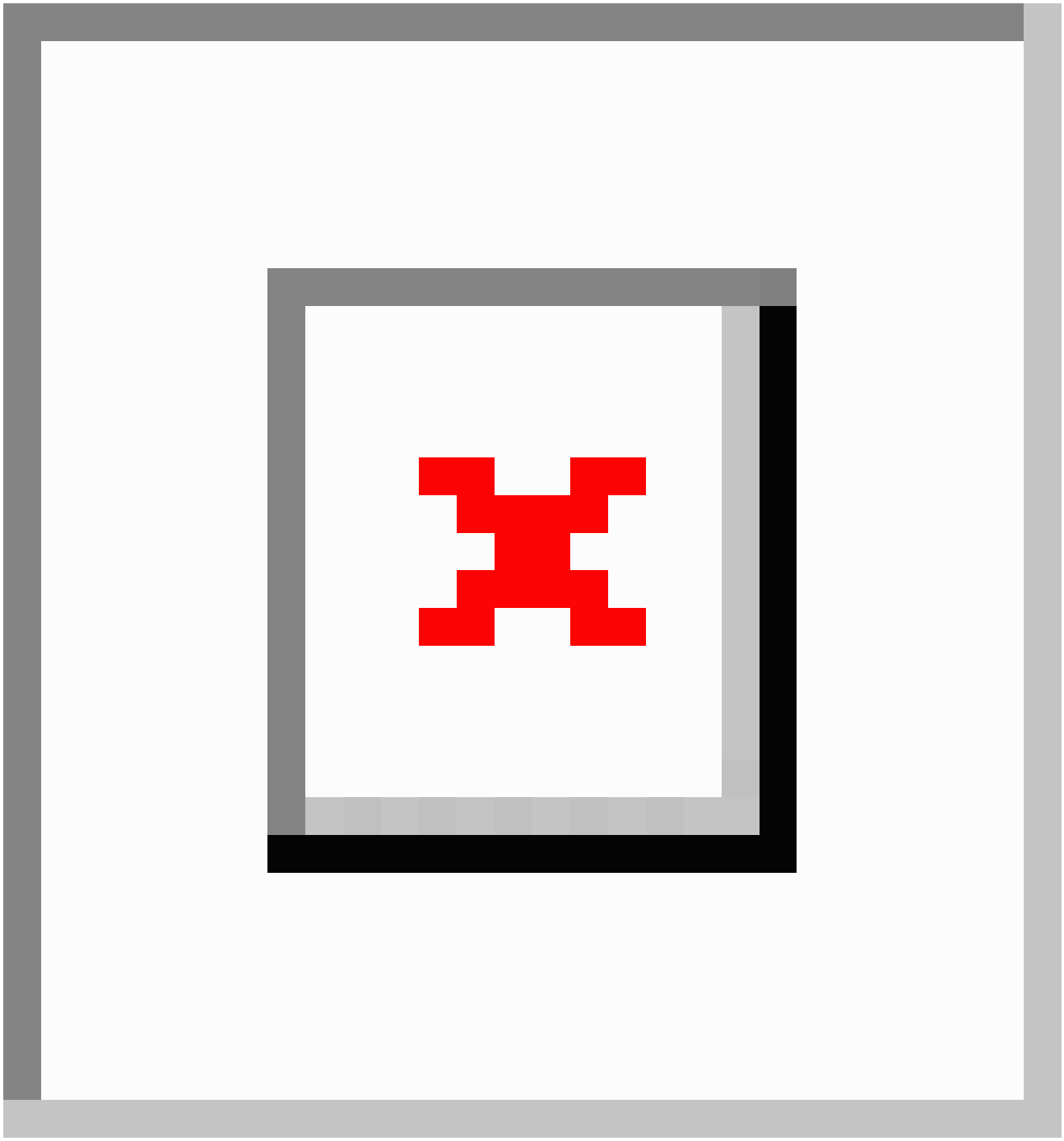
As shown in Figure 1, using data available at different stages, 4 PSNs were generated and named as screening map, echo map, patient map, and surgery map. These data were obtained from the ultrasound reporting system and EHR system of the Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China.

Figure 1. CHDmap contains 4 patient similarity networks generated from 4 different clinical phases, with different data obtained at each phase. CHD: congenital heart disease; ICU: intensive care unit; LOS: length of stay.



A schematic of the data processing and workflow for the construction of the PSN is shown in Figure 2 and described below.

Figure 2. Schematic of data processing and workflow of the construction of the congenital heart disease (CHD) patient similarity network. NLP: natural language processing; t-SNE: t-distributed stochastic neighbor embedding.



Ethical Considerations

This retrospective study was performed according to relevant guidelines and approved by the institutional review board of the Children's Hospital of Zhejiang University School of Medicine with a waiver of informed consent (2018_IRB_078). All cases included in this study were anonymized. Intensive care unit (ICU) clinicians who participated in the trial received cash compensation (RMB ¥100 [US $14.06] per day), which complied with local regulatory requirements for scientific labor.

Data Collection and Preprocessing

In addition to preoperative echocardiography reports that described the CHD conditions, the following patient and surgical characteristics were also collected: age, sex, height, weight, preoperative oxygen saturation of the right-upper limb, surgery time, cardiopulmonary bypass time, aortic cross-clamping time, mechanical ventilation time, duration of postoperative hospital stay, duration of ICU stay, and postoperative complications (the detailed definitions of postoperative complications are shown in Table S1 in Multimedia Appendix 1 [38-40]).

The most challenging part of patient similarity analysis was defining all the semantic concepts in the domain. An ontology of CHD was developed based on reviewing a large number of clinical guidelines for CHD to cover 436 CHD conditions and 87 related echocardiographic indicators. The OWL format ontology file is available on the CHDmap website [41]. The

ontology was used to normalize all concepts and measure semantic similarity among them. It was also used to identify quantitative indicators from the unstructured text of echocardiography reports. In addition to recording some routine cardiac structure indicators, the echocardiography report also provided quantitative indicators regarding various malformations, such as the size of various defects, shunt flow velocity, and pressure difference at the defect, depending on the specific CHD structural malformation. Natural language processing (NLP) technology [38] was used to extract 66 commonly used quantitative indicators. A range of processing and computational methods were used to assess similarity between patients (details information are shown in the supplemental methods and Tables S2 Table S3 in Multimedia Appendix 1). The various automatically extracted measurement values were subject to quality control, and any abnormal data (outside the reasonable range of the corresponding values) were modified or removed after manual verification. The diagnosis in the report was also extracted and mapped to the normalized terms defined in the CHD ontology.

Measuring Patient Similarity

In this study, the similarity of patients with CHD was measured using 4 groups of features: the quantitative echocardiographic indicators, the specific CHD diagnosis, preoperative clinical features, and surgical features. Different distance measurement methods were adopted for different groups of features, as described in the supplemental methods in Multimedia Appendix 1. We provided 3 types of methods to handle the echocardiographic indicators: the origin value, the *z* score, and the indicator combination ratio. The similarity between 2 diagnoses was calculated using the depth of the corresponding nodes in the CHD ontology, which organizes hundreds of CHD diagnoses in a hierarchical structure. Two approaches were used to measure the distance between diagnosis lists: one treats all diagnoses equally, referred to in the result section as “ungrade,” whereas the other distinguishes between basic and other diagnoses, referred to as “grade.” Finally, the patient distance was measured as the weighted sum of the 4 distances as shown in equation (1), and the final distances were also normalized to [0,1].

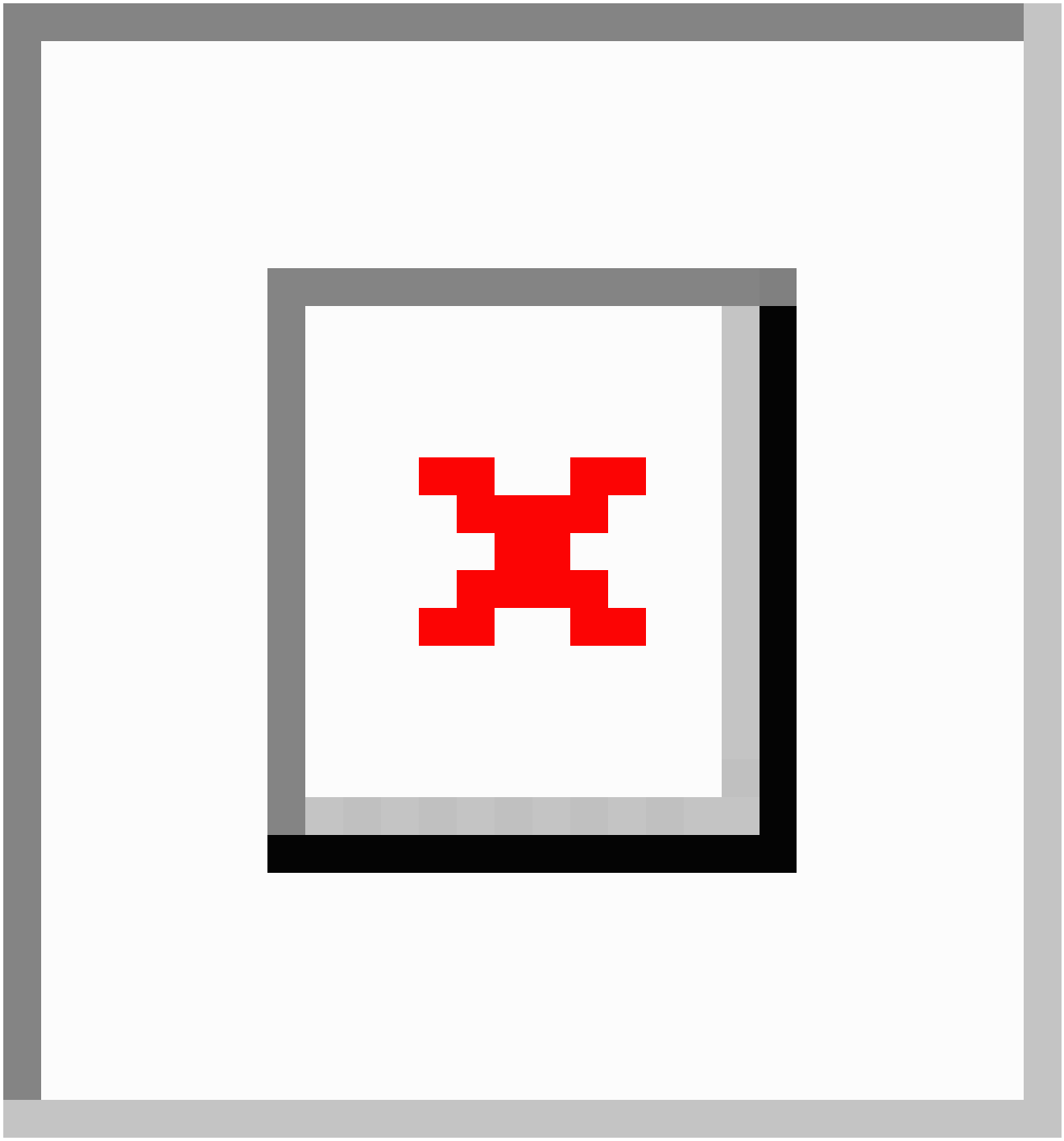
(1)

The weights in equation (1) and the different methods used to measure distance can also be modified by users depending on their experience in different tasks to fully exploit the advanced cognitive ability of clinical professionals. The distance matrix among historical patients can be calculated based on the aforementioned methods. We used t-distributed stochastic neighbor embedding [42] to convert the distance matrix into 2D points, which can be visualized as a map. The user-operable CHDmap was developed based on ECharts [43] using React (Meta) and Node.js (OpenJS Foundation). The patient similarity analysis engine, which measures the distances between a new patient and patients in CHDmap, was developed using Python (Python Software Foundation).

CHDmap

A user-operable CHD PSN called CHDmap was developed and published on the web [44]. The introduction video of this tool is also available in Multimedia Appendix 2. Based on the different available data for each clinical phase, as shown in Figure 1, CHDmap provides 4 different PSNs: the screening map, echo map, patient map, and surgery map. The workspace of CHDmap comprises 3 major modules: (1) map view, (2) cockpit view, and (3) outcome view (as shown in Figure 3).

Figure 3. Screenshot of CHDmap. The map view, cockpit view, and outcome view of the workspace are marked separately. CHDmap was published on the web [44]. CHD: congenital heart disease.



The map view presents the PSN as a zoomable electronic map, in which a node presents a patient and the distance between nodes shows their similarity. The map can be enhanced by using different colors to show the diagnostic labels as well as relevant prognostic indicators (eg, length of stay and complications). Different methods to handle the echocardiographic indicators, such as normal, z score, or combination ratio value, can be selected on the web. The similar patient group is also highlighted on the map view during similarity analysis.

The cockpit view provides a navigation function that helps clinicians locate cases based on specified query conditions, such as age, gender, and CHD subtypes. In practice, clinicians were allowed to create a new case, in which an NLP-based information extraction tool will assist users in filling in most of the echocardiographic indicators based on Chinese echocardiography reports. The top *k* value, or threshold of patient similarity, is used to customize the similar group. For advanced users, a customized map can be generated by adjusting the weights for the patient similarity measurement defined in the *Methods* section.

The outcome view provides an overview of outcomes, including the length of hospital stay, mechanical ventilation time, length of ICU stay, complications, and hospital survival of the selected similar patient group. Multiple charts are used to show the difference between the selected patient group and others. The Mann-Whitney *U* test and the χ^2 test are used to determine the

significance of differences between groups. When there are significant differences between the selected patient group and other patients, the color of the check box at the top of the outcome view will turn red; otherwise, it will stay gray. Checking the box will show detailed charts and tables of the outcome. This real-time feedback will help clinicians adjust the parameters in the cockpit view based on the requirements of the scenario for clinical decision-making. Based on a selected group of similar patients, CHDmap provides machine learning models to personalize the prediction of relevant outcome metrics for the current patient. Therefore, for each case, different parameters can be applied and compared to ultimately assess the credibility of the relevant decision support information.

Evaluation Method

The closer 2 patients are located on the CHDmap, the more similar their conditions and postoperative outcomes are considered to be. When a new patient is admitted to the hospital, historical patients can be divided into similar and nonsimilar groups based on some criteria. There are 2 criteria to define patient similarity groups: one is to use the most similar *k* patients, also known as *k*-nearest neighbor (KNN), to form a patient similarity group, and the other is to define a threshold above which patients form a similarity group. The statistical characteristics or regression value of postoperative outcomes in the similarity group are used to predict the outcomes of the current patient.

In this paper, we evaluated the performance of the surgery map of CHDmap on 2 tasks: predicting postoperative complications as a binary classification task, in which more than 50% of patients in the similarity group with complications were assigned ‘True” for the target patient, and predicting mechanical ventilation duration as a multiple-label classification task (I: 0-12 h, II: 12-24 h, III: 24-48 h, and IV: >48 h), in which the category with the highest proportion in the similarity group was assigned to the target patient.

As the optimum *k* of KNN to form a similarity group for a specific case is always different, the unified population-level optimized *k* on the training data set was used to evaluate CHDmap on the test data set without individual customization. Different data preprocessing methods (original, *z* score, and combination ratio) and whether to distinguish primary diagnoses (grade and ungrade) were tested and compared.

Making decisions may not be straightforward if the outcome of a similar patient group is extremely heterogeneous, whereby a machine learning model based on a similar patient population can provide a more personalized prediction of the relevant prognostic indicators. Although there are numerous machine learning models to choose from, the focus of this study was to demonstrate the advantages of basing the model on similar patient populations, so we chose to use the most conventional and easily understood logistic regression (LR) model. Clinical users obtained a population of similar patients after various parameter adjustments and threshold settings on CHDmap, and the data from this population were used to train an LR model (KNN+LR), which can be accomplished on the web in real time because this population of similar patients is usually not very large. To demonstrate the effect of similar patient populations, we trained another LR model (*k*-Random+LR) based on randomly collected cases of the same size in parallel in the evaluation. We evaluated such approaches and compared the LR models based on *k* similar patients and *k* random patients.

The accuracy, recall, F_1-score, and area under the receiver operating characteristic curve (AUC), which are defined below, were adopted to evaluate the performance of the classification. Accuracy is defined as the total correctly classified example including true positive (TP) and true negative (TN) divided by the total number of classified examples. Recall quantifies the number of correct positive predictions made out of all positive predictions that could have been made. F_1-score is a weighted average of precision and recall. As we know, in precision and recall, there are false positive (FP) and false negative (FN), so F_1-score also considers both of them. AUC provides an aggregate measure of the performance across all possible classification thresholds. The higher the accuracy, recall, F_1-score, and AUC, the better the model’s performance is at distinguishing between the positive and negative classes.

(2)

(3)

(4)

(5)

The performance was evaluated on an independent test set, which included 256 patients with CHD. These test cases were also available on CHDmap when users created a new case. Three clinicians working in the cardiac ICU with extensive experience were also asked to make relevant judgments for these test cases based on their clinical experience. After half a year following the initial trial, we conducted an experiment where the 3 clinicians were asked to make further predictions based on the output of CHDmap, and this prediction was compared with the previous results based on clinical experience alone to validate the benefits of CHDmap in supporting clinical decision-making.

Results

Population Characteristics

A total of 4774 patients who underwent congenital heart surgery between June 2016 and June 2021 at the Children’s Hospital of Zhejiang University School of Medicine were used to generate the CHD PSN. The performance of the PSN in predicting complications and mechanical ventilation duration was evaluated on an independent test data set, which included 256 pediatric patients who underwent congenital heart surgery between July 2021 and November 2021 at the Children’s Hospital of Zhejiang University School of Medicine. The characteristics of patients used to generate the PSN and for evaluation are described in Table 1. Since the test data and the

data used by the PSN were generated and collected in different time periods, as shown in Table 1, they are somewhat statistically different. The test data were older; therefore, the patients were significantly larger in terms of height and weight (P<.001), and there were also relatively large differences in the distribution of outcomes, lower complication rates, and shorter duration of mechanical ventilation. It should be noted that the diagnostic label is not the complete diagnostic information; we just use a few of the most common CHD subtypes to facilitate statistics and visualization, and this cohort contains a complete range of epidemiological characteristics as well as a variety of complex CHD subtypes such as transposition of the great arteries, tetralogy of Fallot, etc, which may appear in various diagnostic labels that they are combined with. When the case has 2 common CHD subtypes, such as ventricular septal defect and patent ductus arteriosus, only the more common subtype, ventricular septal defect, is labeled.

Table . Characteristics of patients with CHD[a] used to generate CHDmap and in the test data set.

Characteristic		Patients of CHDmap (n=4774)	Patients of the test data set (n=256)	*P* value
Gender (male), n (%)		2336 (48.9)	111 (43.4)	.09
Age (mo), median (IQR)		12.0 (4.0-32.0)	22.1 (7.8-50.9)	<.001
Height (cm), median (IQR)		75.0 (63.0-94.0)	85.5 (67.0-106.3)	<.001
Weight (kg), median (IQR)		9.2 (6.0-13.4)	10.8 (6.8-16.5)	<.001
Preoperative oxygen saturation (%), median (IQR)		98.0 (97.0-99.0)	98.0 (97.0-99.0)	.007
Surgery time (min), median (IQR)		119.0 (96.0-147.0)	120.0 (100.0-147.0)	.25
Cardiopulmonary bypass time (min), median (IQR)		60.0 (48.0-82.0)	61.5 (49.3-80.0)	.55
Aortic cross-clamping time (min), median (IQR)		40.0 (28.0-54.0)	38.5 (27.0-52.0)	.55
Duration of hospital stay (d), median (IQR)		9.0 (7.0-13.0)	7.0 (6.0-11.0)	.003
Duration of ICU[b] stay (d), median (IQR)		3.0 (1.0-4.0)	3.0 (1.0-4.0)	.49
Diagnostic label, n (%)				.46
	ASD[c] and VSD[d]	1659 (34.8)	78 (30.5)	
	VSD	1522 (31.9)	94 (36.7)	
	ASD	1228 (25.7)	65 (25.4)	
	PFO[e]	134 (2.8)	5 (2)	
	PDA[f]	123 (2.6)	9 (3.5)	
	Others	108 (2.3)	5 (2)	
Mechanical ventilation time (%), n (%)				.001
	I (<12 h)	3009 (63.0)	180 (70.3)	
	II (12-24 h)	918 (19.2)	54 (21.1)	
	III (24-48 h)	433 (9.1)	7 (2.7)	
	IV (≥48 h)	414 (8.7)	15 (5.9)	
Complication, n (%)		1229 (25.7)	48 (18.8)	.02

[a]CHD: congenital heart disease.

[b]ICU: intensive care unit.

[c]ASD: atrial septal defect.

[d]VSD: ventricular septal defect.

[e]PFO: patent foramen ovale.

[f]PDA: patent ductus arteriosus.

Performance of CHDmap

Three methods for preprocessing the echocardiographic indicators (origin, *z* score, combination) and 2 distinguishing primary diagnoses (grade and ungrade) were used to compare their effect on CHDmap performance. The performance of the CHDmap and 3 clinicians is shown in Table 2 and Figure 4.

Table . Evaluation results in the 2 tasks.

Methods		Prediction of postoperative complications				Prediction of mechanical ventilation duration			
		Accuracy	Recall	F_1-score	AUC[a]	Accuracy	Recall	F_1-score	AUC
KNN[b]									
	Origin+ungrade	0.832	0.438	0.494	0.757	0.813	0.444	0.459	0.862
	Origin+grade	0.836	0.417	0.489	0.773	0.797	0.437	0.467	0.860
	z score+ungrade	0.828	0.458	0.500	0.738	0.836	0.554	0.574	0.902
	z score+grade	0.848	0.458	0.530	0.747	0.855	0.564	0.573	0.895
	Combination+ungrade	0.836	0.500	0.533	0.767	0.828	0.468	0.488	0.900
	Combination+grade	0.859	0.458	0.550	0.768	0.855	0.521	0.545	0.873
KNN+LR[c]									
	Origin+ungrade	0.813	0.604	0.547	*0.810*[d]	0.848	0.558	0.602	0.921
	Origin+grade	0.813	*0.667*	0.571	0.799	0.863	0.589	0.632	0.920
	z score+ungrade	0.809	0.604	0.542	0.809	0.840	0.537	0.561	0.888
	z score+grade	0.813	0.646	0.564	0.805	0.855	0.549	0.562	0.886
	Combination+ungrade	0.805	0.583	0.528	0.801	0.840	0.537	0.555	0.900
	Combination+grade	0.805	0.604	0.537	0.798	0.824	0.500	0.522	*0.926*
k-Random+LR		0.809	0.500	0.495	0.774	0.809	0.484	0.488	0.895
Clinicians[e]									
	C1	0.875	0.396	0.543	N/A[f]	0.844	*0.614*	0.618	N/A
	C2	0.758	0.646	0.500	N/A	0.734	0.535	0.496	N/A
	C3	0.840	0.208	0.328	N/A	0.797	0.498	0.536	N/A
	Clinician average	0.824	0.417	0.457	N/A	0.792	0.549	0.550	N/A
	C1+CHDmap	*0.883*	0.426	*0.580*	N/A	*0.943*	0.612	*0.647*	N/A
	C2+CHDmap	0.816	0.5625	0.534	N/A	0.874	0.587	0.542	N/A
	C3+CHDmap	0.852	0.313	0.441	N/A	0.916	0.511	0.546	N/A
	Clinician+CHDmap average	0.850	0.434	0.518	N/A	0.911	0.570	0.578	N/A

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

[b]KNN: *k*-nearest neighbor.

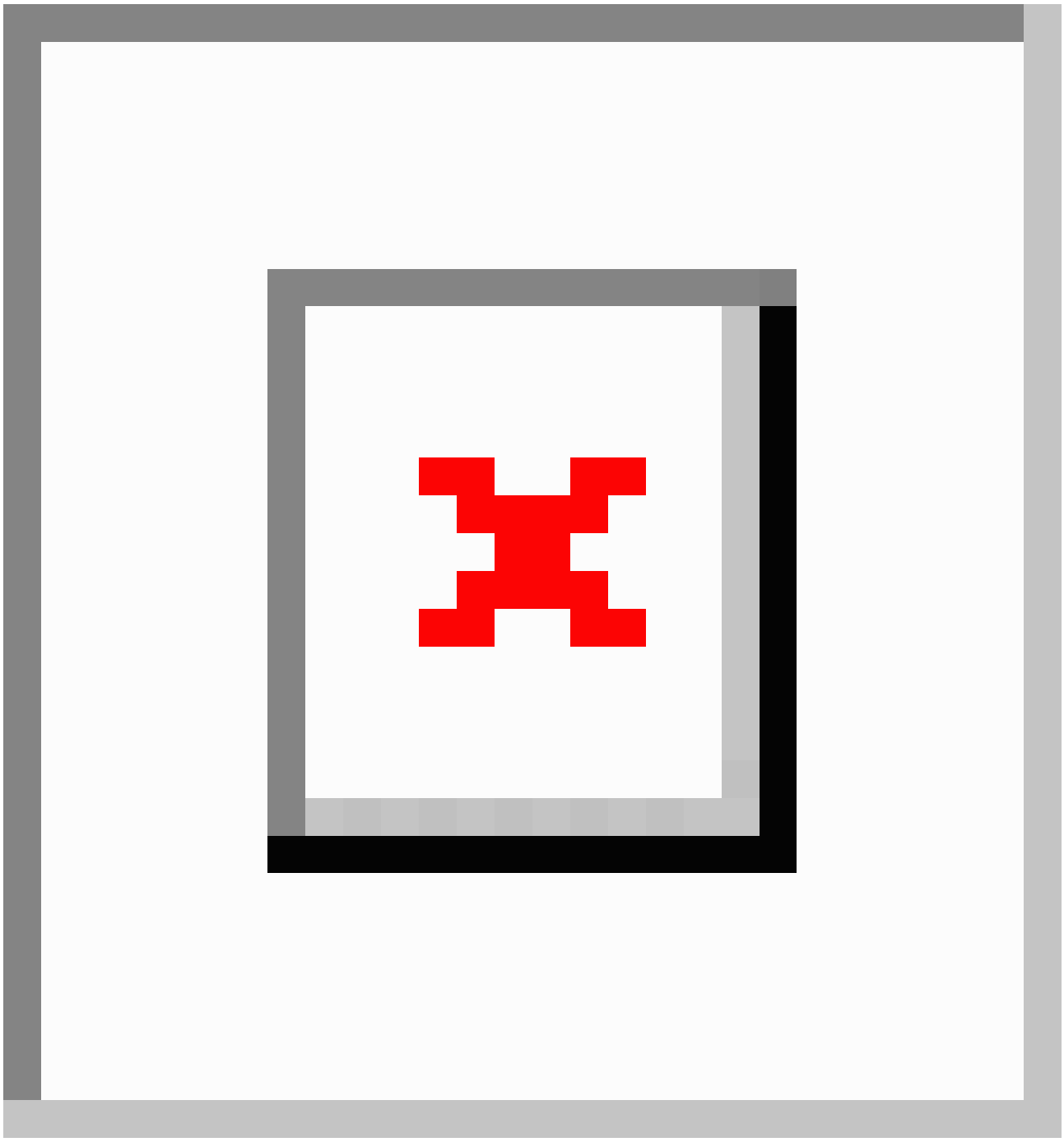
[c]LR: logistic regression.

[d]In each column, the maximum value is italicized.

[e]The performance of the 3 clinicians are labeled as C1, C2, and C3.

[f]N/A: not applicable.

Figure 4. Evaluation result based on receiver operating characteristic curves. (A) Binary postoperative complication prediction using KNN; (B) to (E) multilabel mechanical ventilation duration prediction (I: 0-12 h, II: 12-24 h, III: 24-48 h, and IV: >48 h) using KNN, respectively; (F) binary postoperative complication prediction using KNN+LR; (G) to (J) multilabel mechanical ventilation duration prediction (I: 0-12 h, II: 12-24 h, III: 24-48 h, and IV: >48 h) using KNN+LR, respectively. The performance of 3 clinicians are labeled as black stars in different tasks as C1, C2, and C3. The performance of 3 clinicians enhanced by CHDmap are labeled as red stars. CHD: congenital heart disease; KNN: *k*-nearest neighbor; LR: logistic regression.



In the postoperative complication prediction task, the F_1-score of methods using KNN exceeded the average of the 3 clinicians, although 1 clinician achieved the best accuracy when dropping a high recall value. In all 6 KNN methods, introducing the indicator combination ratio and distinguishing the primary diagnosis in the similarity measurement can truly improve the overall performance of the F_1-score. LR models constructed using the KNN-obtained patient groups were able to generally achieve better predictions compared to simple voting of similar patients and the LR model based on k random patients. Interestingly, both the model with the best F_1-score performance and the model with the best AUC used the original values. This may be because original values are more reflective of individualized patient differences in a similar patient population. The main improvement of CHDmap on this task is reflected in the general improvement in recall values, with the best recall method being 0.250 higher than the clinician average.

In another multiclassification task that predicts mechanical ventilation duration, the differences among these different KNN methods in overall performance were not consistent. The

KNN+LR approaches also achieved better composite performance (F_1-score and AUC), although 1 of the human experts got the best recall value.

From the test result, clinicians do not have the same performance for such predictive judgments. Some raise the standard and thus miss some events; on the other hand, some lower the judgment threshold, and thus the accuracy of the judgment decreases. At the same time, the performance of clinical experts on different tasks is inconsistent. A simple poll of the *k*-most similar patients provided by the CHDmap can achieve better results than the clinician average. When 3 clinicians were allowed to use the results of CHDmap (KNN+LR) as a reference to give predictions again, all 3 clinicians achieved a substantial improvement in their prediction ability. The averages of accuracy, recall, and F_1-score in the first task improved by 0.026, 0.017, and 0.061, respectively. The averages of accuracy, recall, and F_1-score in the second task improved by 0.119, 0.021, and 0.028, respectively. One of the enhanced clinicians also surpassed the KNN+LR CHDmap.

It is important to note that the evaluation is performed with population-optimized parameters, whereas in practice, clinicians can adjust the relevant parameters such as *k* or similarity threshold for each case in a personalized manner, which theoretically leads to better results. The use of the obtained similar patient population to construct modern deep learning models for prediction can further improve the performance of each prediction task. Especially important is that the experience and cognitive ability of the clinical expert combined with CHDmap can further enhance the accuracy of the prediction.

Discussion

Principal Findings

Medicine remains both an art and a science, which are congruent to the extent that the individual patient resembles the average subject in randomized controlled trials. Although the evidence-based medicine approach proposes personalized care, it still fails to address the physician's most important question—"How to treat the unique patient in front of me?"—in many real clinical scenarios where the complexity of the situation makes none of the available evidence applicable [45]. The proposal of MBE represents a fundamental change in clinical decision-making [5,6]. Although how to construct an MBE clinical decision support tool still faces many challenges, the CHDmap seems to be a very promising first step in realizing what has been coined MBE.

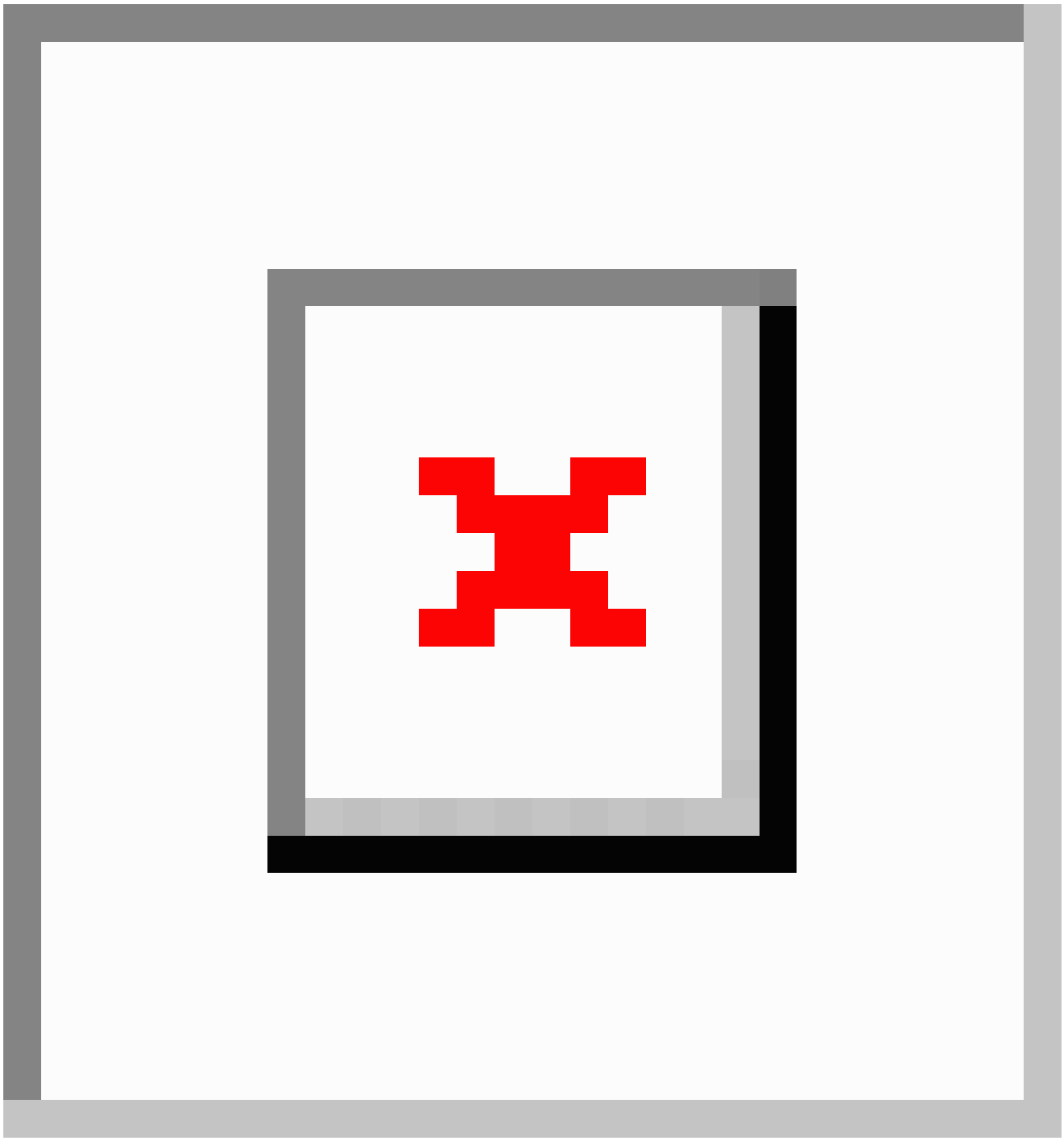
AI is poised to reshape health care. Many AI applications, especially modern deep learning models, have been developed in recent years to improve clinical prediction abilities. In addition to supervised and unsupervised machine learning, PSNs, another form of data-driven AI, have shown many unique properties in the clinical field, especially in complex clinical settings such as surgery for CHD. Moreover, their potential to construct a "library of clinical experience" will gradually be recognized, discovered, and used in the context of the continuous accumulation of medical big data.

In many other popular AI paradigms, such as supervised or unsupervised machine learning, models are usually trained toward a specific task, and thus, the models are only capable of performing that single task. This, coupled with the black-box nature of many machine learning models, especially deep learning models, makes it difficult to widely apply these techniques in practice. In contrast, patient similarity analysis exhibits many natural advantages. First, PSNs usually do not serve a single task; all characteristics exhibited by the patient similarity group, such as disease risk, various prognostic outcomes, and cost of care, can be used as MBE for decision support. Second, instead of a model that simply gives black-box predictions, CHDmap allows users to see how the patient similarity group is segmented and bounded across the patient population and then adjust the size of the patient similarity group or set custom quantitative thresholds based on their knowledge and experience. On CHDmap, the results after parameter adjustments during user manipulation are reflected in the visualized map in real time, and the statistical characteristics of multiple predictors that distinguish the current patient's similar group from other patients are also highlighted by the color of the title of the outcome view. The process of continuously adjusting and optimizing parameters through visualized feedback combines the computational advantages of computers and the advanced cognitive abilities of the human brain and truly puts the clinician, who is responsible for the decision, in control of the decision-making. Third, many machine learning models tend to require that the test and training data have consistent statistical distribution characteristics, but as shown in this evaluation, similarity analyses are still very compatible with test data with different characteristics. Finally, this PSN framework does not exclude any type of machine learning models, and all models constructed based on similar patient populations are expected to be more adaptable to individualized decision-making needs than models trained on heterogeneous populations.

Because the goal of patient similarity analysis is to be able to mimic clinical analogy reasoning, the major challenge is constructing computational patient similarity measurements that are consistent with sophisticated clinical reasoning. This is especially true when faced with complex scenarios containing a large number of dynamic features with different dimensions. Some deep learning models have been introduced to address this challenge [46-49], but they do not exhibit the interpretability and tractability of PSNs. Another way to address this challenge is to open up the computational process to clinicians, allowing them to determine and adjust the weights of different dimensions and thresholds for the similarity group themselves, thus better simulating their clinical reasoning process, as shown in Figure 5. We believe that clinical users will be able to learn how to better optimize these parameters as they continue to gain experience and understanding of this "large history data set" in the process of using CHDmap. Using a data-driven approach on how to customize the parameters of PSNs to be able to self-optimize and adapt to different tasks is also a good research direction for the future. In this study, CHDmap serves as a personalized decision aid for clinicians, using the computer's power in data storage and processing while giving clinicians more control over the decision-making process. We believe

CHDmap can perform better with the full involvement of clinicians.

Figure 5. Collaborative decision-making based on the congenital heart disease patient similarity network (PSN). The right half shows the storage and computational capacity of the PSN for a large number of cases; the left half shows the role of the clinical user who, by receiving a variety of feedback and his or her own experience, can autonomously adjust the parameters of the similarity group and reconstruct the similarity network so that the strengths of both can be used to make collaborative decisions. ASD: atrial septal defect; PDA: patent ductus arteriosus; PFO: patent foramen ovale; VSD: ventricular septal defect.



CHDmap can be used in several scenarios: for the intensivists in cardiac ICUs, CHDmap can be used to predict postoperative complications after cardiac surgery, as evaluated in this paper; for surgeons, CHDmap can also be used to assess the prognosis of surgical procedures; and for departmental managers, CHDmap can be used to assess the lengths of stay and costs. By far, CHDmap is still in the early stages of a research project. Transforming this tool into routine care is dependent on the availability of funding and the willingness of users to change their existing working patterns. The publication of this paper will also facilitate the advancement of our subsequent translational work.

It is important to note that associations between treatments and outcomes obtained by observation in similar patient populations may not be causal. The real causal effects often rely on a matching process to control for the bias introduced by the treatment itself in the selection of patients [50]. An initial demo feature is available on CHDmap to estimate treatment outcome

effects based on matched patient groups. CHDmap can match 1 or *k* patients for each patient receiving the treatment using a PSN and then allow for a more visual and unbiased assessment of treatment outcomes by showing the difference in prognosis between these 2 groups of patients. It is important to note that this causal assessment assumes that there are no other factors outside the variables covered by the patient's similarity analysis that may influence treatment choice or prognosis. Thus, the reliability of this real world–generated evidence usually relies on clinical experts to judge it as well. In future versions, we hope to incorporate more modern frameworks for causal inference (such as DoWhy [51]) to automatically quantitatively assess causal effects as well as their reliability.

There are several limitations to this study. First, limited clinical features were used to measure the similarity of patients with CHD. In addition to the information presented by the echocardiography, there is a wealth of other clinical information that can be used to assess the patient's status. Second, the use of NLP to automatically extract measurement information can also be subject to errors or mismatches, and although manual quality control is carried out, it is still not possible to ensure that all of the measurements are 100% accurate. Third, just as clinicians gain clinical experience by continuously treating different patients, PSNs need to expand their ability to dynamically accumulate cases. A PSN with a web-based automatic update mechanism will be the next key research step. Fourth, data from only a single center were used to evaluate this tool, and the introduction of data from multiple centers during PSN construction may pose unknown risks that require attention in future studies. Finally, different clinicians may have different decision-making philosophies, and different weights can be assigned to different indicators for different tasks. CHDmap offers only a limited number of customizations that may be difficult to adapt to all scenarios. A way to attribute weights to each of the indicators and dimensions by AI for specific tasks may potentially improve the performance of CHDmap in the future.

Conclusions

A clinician-operable PSN for CHD was proposed and developed to help clinicians make decisions based on thousands of previous surgery cases. Without individual optimization, CHDmap can obtain competitive performance compared to clinical experts. Statistical analysis of data based on patient similarity groups is intuitive and clear to clinicians, whereas the operable, visual user interface puts clinicians in real control of decision-making. Clinicians supported by CHDmap can make better decisions than both pure experience-based decisions and AI model output results. Such a PSN-based framework can become a routine method of CHD case management and use. The MBE can be embraced in clinical practice, and its full potential can be realized.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81871456).

Authors' Contributions

HL, SS, and QS contributed equally to the paper as cocorresponding authors. SS can be contacted at Sicu1@zju.edu.cn, and QS can be contacted at shuqiang@zju.edu.cn.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental methods, definitions of postoperative complications, features used to measure patient similarity, and echocardiographic indicators used in different calculations.
[DOCX File, 1306 KB - medinform_v12i1e49138_app1.docx]

Multimedia Appendix 2
Video introduction for CHDmap.
[MP4 File, 102353 KB - medinform_v12i1e49138_app2.mp4]

References

1. Bernier PL, Stefanescu A, Samoukovic G, Tchervenkov CI. The challenge of congenital heart disease worldwide: epidemiologic and demographic facts. Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu 2010;13(1):26-34. [doi: 10.1053/j.pcsu.2010.02.005] [Medline: 20307858]
2. van der Linde D, Konings EEM, Slager MA, et al. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. J Am Coll Cardiol 2011 Nov 15;58(21):2241-2247. [doi: 10.1016/j.jacc.2011.08.025] [Medline: 22078432]

3. Jacobs JP, Mayer JEJ, Mavroudis C, et al. The Society of Thoracic Surgeons Congenital Heart Surgery Database: 2016 update on outcomes and quality. Ann Thorac Surg 2016 Mar;101(3):850-862. [doi: 10.1016/j.athoracsur.2016.01.057] [Medline: 26897186]
4. Triedman JK, Newburger JW. Trends in congenital heart disease. Circulation 2016 Jun 21;133(25):2716-2733. [doi: 10.1161/CIRCULATIONAHA.116.023544] [Medline: 27324366]
5. Horwitz RI, Hayes-Conroy A, Caricchio R, Singer BH. From evidence based medicine to medicine based evidence. Am J Med 2017 Nov;130(11):1246-1250. [doi: 10.1016/j.amjmed.2017.06.012] [Medline: 28711551]
6. van den Eynde J, Manlhiot C, van de Bruaene A, et al. Medicine-based evidence in congenital heart disease: how artificial intelligence can guide treatment decisions for individual patients. Front Cardiovasc Med 2021 Dec;8:798215. [doi: 10.3389/fcvm.2021.798215] [Medline: 34926630]
7. Benavidez OJ, Gauvreau K, del Nido P, Bacha E, Jenkins KJ. Complications and risk factors for mortality during congenital heart surgery admissions. Ann Thorac Surg 2007 Jul;84(1):147-155. [doi: 10.1016/j.athoracsur.2007.02.048] [Medline: 17588402]
8. Pasquali SK, He X, Jacobs JP, Jacobs ML, O'Brien SM, Gaynor JW. Evaluation of failure to rescue as a quality metric in pediatric heart surgery: an analysis of the STS Congenital Heart Surgery Database. Ann Thorac Surg 2012 Aug;94(2):573-580. [doi: 10.1016/j.athoracsur.2012.03.065] [Medline: 22633496]
9. Kansy A, Tobota Z, Maruszewski P, Maruszewski B. Analysis of 14,843 neonatal congenital heart surgical procedures in the European Association for Cardiothoracic Surgery Congenital Database. Ann Thorac Surg 2010 Apr;89(4):1255-1259. [doi: 10.1016/j.athoracsur.2010.01.003] [Medline: 20338347]
10. Jenkins KJ, Gauvreau K, Newburger JW, Spray TL, Moller JH, Iezzoni LI. Consensus-based method for risk adjustment for surgery for congenital heart disease. J Thorac Cardiovasc Surg 2002 Jan;123(1):110-118. [doi: 10.1067/mtc.2002.119064] [Medline: 11782764]
11. Lacour-Gayet F, Clarke D, Jacobs J, et al. The Aristotle score: a complexity-adjusted method to evaluate surgical results1. Eur J Cardiothorac Surg 2004 Jun;25(6):911-924. [doi: 10.1016/j.ejcts.2004.03.027] [Medline: 15144988]
12. O'Brien SM, Clarke DR, Jacobs JP, et al. An empirically based tool for analyzing mortality associated with congenital heart surgery. J Thorac Cardiovasc Surg 2009 Nov;138(5):1139-1153. [doi: 10.1016/j.jtcvs.2009.03.071] [Medline: 19837218]
13. Jacobs ML, O'Brien SM, Jacobs JP, et al. An empirically based tool for analyzing morbidity associated with operations for congenital heart disease. J Thorac Cardiovasc Surg 2013 Apr;145(4):1046-1057.E1. [doi: 10.1016/j.jtcvs.2012.06.029] [Medline: 22835225]
14. Kalfa D, Krishnamurthy G, Duchon J, et al. Outcomes of cardiac surgery in patients weighing <2.5 kg: affect of patient-dependent and -independent variables. J Thorac Cardiovasc Surg 2014 Dec;148(6):2499-2506.E1. [doi: 10.1016/j.jtcvs.2014.07.031] [Medline: 25156464]
15. Agarwal HS, Wolfram KB, Saville BR, Donahue BS, Bichell DP. Postoperative complications and association with outcomes in pediatric cardiac surgery. J Thorac Cardiovasc Surg 2014 Aug;148(2):609-616.E1. [doi: 10.1016/j.jtcvs.2013.10.031] [Medline: 24280709]
16. Zeng X, An J, Lin R, et al. Prediction of complications after paediatric cardiac surgery. Eur J Cardiothorac Surg 2020 Feb 1;57(2):350-358. [doi: 10.1093/ejcts/ezz198] [Medline: 31280308]
17. Zeng X, Hu Y, Shu L, et al. Explainable machine-learning predictions for complications after pediatric congenital heart surgery. Sci Rep 2021 Aug 26;11(1):17244. [doi: 10.1038/s41598-021-96721-w] [Medline: 34446783]
18. Zeng X, Shi S, Sun Y, et al. A time-aware attention model for prediction of acute kidney injury after pediatric cardiac surgery. J Am Med Inform Assoc 2022 Dec 13;30(1):94-102. [doi: 10.1093/jamia/ocac202] [Medline: 36287639]
19. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
20. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67. [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
21. Hu Y, Gong X, Shu L, et al. Understanding risk factors for postoperative mortality in neonates based on explainable machine learning technology. J Pediatr Surg 2021 Dec;56(12):2165-2171. [doi: 10.1016/j.jpedsurg.2021.03.057] [Medline: 33863558]
22. Pai S, Bader GD. Patient similarity networks for precision medicine. J Mol Biol 2018 Sep 14;430(18 Pt A):2924-2938. [doi: 10.1016/j.jmb.2018.05.037] [Medline: 29860027]
23. Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: a systematic review. J Biomed Inform 2018 Jul;83:87-96. [doi: 10.1016/j.jbi.2018.06.001] [Medline: 29864490]
24. Zeng X, Jia Z, He Z, et al. Measure clinical drug–drug similarity using electronic medical records. Int J Med Inform 2019 Apr;124:97-103. [doi: 10.1016/j.ijmedinf.2019.02.003] [Medline: 30784433]
25. Jia Z, Lu X, Duan H, Li H. Using the distance between sets of hierarchical taxonomic clinical concepts to measure patient similarity. BMC Med Inform Decis Mak 2019 Apr 25;19(1):91. [doi: 10.1186/s12911-019-0807-y] [Medline: 31023325]
26. Cheng F, Liu D, Du F, et al. VBridge: connecting the dots between features and data to explain healthcare models. IEEE Trans Vis Comput Graph 2022 Jan;28(1):378-388. [doi: 10.1109/TVCG.2021.3114836] [Medline: 34596543]

27. Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: interpretable patient classification using integrated patient similarity networks. Mol Syst Biol 2019 Mar 14;15(3):e8497. [doi: 10.15252/msb.20188497] [Medline: 30872331]
28. Yang J, Dong C, Duan H, Shu Q, Li H. RDmap: a map for exploring rare diseases. Orphanet J Rare Dis 2021 Feb 25;16(1):101. [doi: 10.1186/s13023-021-01741-4] [Medline: 33632281]
29. Zhang G, Peng Z, Yan C, Wang J, Luo J, Luo H. A novel liver cancer diagnosis method based on patient similarity network and DenseGCN. Sci Rep 2022 Apr 26;12(1):6797. [doi: 10.1038/s41598-022-10441-3] [Medline: 35474072]
30. Jia Z, Zeng X, Duan H, Lu X, Li H. A patient-similarity-based model for diagnostic prediction. Int J Med Inform 2020 Mar;135:104073. [doi: 10.1016/j.ijmedinf.2019.104073] [Medline: 31923816]
31. Li L, Cheng WY, Glicksberg BS, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med 2015 Oct 28;7(311):311ra174. [doi: 10.1126/scitranslmed.aaa9364] [Medline: 26511511]
32. Tokodi M, Shrestha S, Bianco C, et al. Interpatient similarities in cardiac function: a platform for personalized cardiovascular medicine. JACC Cardiovasc Imaging 2020 May;13(5):1119-1132. [doi: 10.1016/j.jcmg.2019.12.018] [Medline: 32199835]
33. Wang N, Wang M, Zhou Y, et al. Sequential data-based patient similarity framework for patient outcome prediction: algorithm development. J Med Internet Res 2022 Jan 6;24(1):e30720. [doi: 10.2196/30720] [Medline: 34989682]
34. 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]
35. Tan WY, Gao Q, Oei RW, Hsu W, Lee ML, Tan NC. Diabetes medication recommendation system using patient similarity analytics. Sci Rep 2022 Dec 3;12(1):20910. [doi: 10.1038/s41598-022-24494-x] [Medline: 36463296]
36. Chen X, Faviez C, Vincent M, et al. Patient-patient similarity-based screening of a clinical data warehouse to support ciliopathy diagnosis. Front Pharmacol 2022 Mar 25;13:786710. [doi: 10.3389/fphar.2022.786710] [Medline: 35401179]
37. Gentner D. Structure‐mapping: a theoretical framework for analogy. Cognitive Science 1983;7(2):155-170 [FREE Full text]
38. Shi Y, Li Z, Jia Z, et al. Automatic knowledge extraction and data mining from echo reports of pediatric heart disease: application on clinical decision support. In: Sun M, Liu Z, Zhang M, Liu Y, editors. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. CCL 2015, NLP-NABD 2015. Lecture Notes in Computer Science, vol 9427: Springer; 2015:417-424. [doi: 10.1007/978-3-319-25816-4_34]
39. Lopez L, Colan S, Stylianou M, et al. Relationship of echocardiographic z scores adjusted for body surface area to age, sex, race, and ethnicity: the Pediatric Heart Network Normal Echocardiogram Database. Circ Cardiovasc Imaging 2017 Nov;10(11):e006979. [doi: 10.1161/CIRCIMAGING.117.006979] [Medline: 29138232]
40. Zhou M, Yu J, Duan H, et al. Study on the correlation between preoperative echocardiography indicators and postoperative prognosis in children with ventricular septal defect. Article in Chinese. Chinese J Ultrason 2022 Sep 25;31(9):767-773. [doi: 10.3760/cma.j.cn131148-20220127-00076]
41. Download. CHDmap. URL: http://chdmap.nbscn.org/Help#download [accessed 2024-01-04]
42. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(86):2579-2605 [FREE Full text]
43. Li D, Mei H, Shen Y, et al. ECharts: a declarative framework for rapid construction of web-based visualization. Vis Inform 2018 Jun;2(2):136-146. [doi: 10.1016/j.visinf.2018.04.011]
44. CHDmap. URL: http://chdmap.nbscn.org/ [accessed 2024-01-04]
45. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. NPJ Digit Med 2020 May 11;3:67. [doi: 10.1038/s41746-020-0277-8] [Medline: 32411828]
46. Suo Q, Ma F, Yuan Y, et al. Deep patient similarity learning for personalized healthcare. IEEE Trans Nanobioscience 2018 Jul;17(3):219-227. [doi: 10.1109/TNB.2018.2837622] [Medline: 29994534]
47. Gu Y, Yang X, Tian L, et al. Structure-aware Siamese graph neural networks for encounter-level patient similarity learning. J Biomed Inform 2022 Mar;127:104027. [doi: 10.1016/j.jbi.2022.104027] [Medline: 35181493]
48. Sun Z, Lu X, Duan H, Li H. Deep dynamic patient similarity analysis: model development and validation in ICU. Comput Methods Programs Biomed 2022 Oct;225:107033. [doi: 10.1016/j.cmpb.2022.107033] [Medline: 35905698]
49. Navaz AN, El-Kassabi HT, Serhani MA, Oulhaj A, Khalil K. A novel patient similarity network (PSN) framework based on multi-model deep learning for precision medicine. J Pers Med 2022 May 10;12(5):768. [doi: 10.3390/jpm12050768] [Medline: 35629190]
50. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010 Feb 1;25(1):1-21. [doi: 10.1214/09-STS313] [Medline: 20871802]
51. Sharma A, Syrgkanis V, Zhang C, Kıcıman E. DoWhy: addressing challenges in expressing and validating causal assumptions. arXiv. Preprint posted online on Aug 27, 2021. [doi: 10.48550/arXiv.2108.13518]

Abbreviations

AI: artificial intelligence
AUC: area under the receiver operating characteristic curve
CHD: congenital heart disease

EHR: electronic health record
FN: false negative
FP: false positive
ICU: intensive care unit
KNN: *k*-nearest neighbor
LR: logistic regression
MBE: medicine-based evidence
NLP: natural language processing
PSN: patient similarity network
RACHS-1: Risk Adjustment for Congenital Heart Surgery 1
STS-EACTS: Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery
TN: true negative
TP: true positive

Edited by C Lovis; submitted 24.05.23; peer-reviewed by JVD Eynde, Y Kim; revised version received 21.08.23; accepted 16.11.23; published 19.01.24.

Please cite as:
Li H, Zhou M, Sun Y, Yang J, Zeng X, Qiu Y, Xia Y, Zheng Z, Yu J, Feng Y, Shi Z, Huang T, Tan L, Lin R, Li J, Fan X, Ye J, Duan H, Shi S, Shu Q
A Patient Similarity Network (CHDmap) to Predict Outcomes After Congenital Heart Surgery: Development and Validation Study
JMIR Med Inform 2024;12:e49138
URL: https://medinform.jmir.org/2024/1/e49138
doi:10.2196/49138

Original Paper

Prediction of Antibiotic Resistance in Patients With a Urinary Tract Infection: Algorithm Development and Validation

Nevruz İlhanlı[1,2*], MSc; Se Yoon Park[1,3,4*], MD, PhD; Jaewoong Kim[1,3], MSc; Jee An Ryu[1], BA; Ahmet Yardımcı[2], PhD; Dukyong Yoon[1,4,5], MD, PhD

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Yongin, Republic of Korea

[2]Department of Biostatistics and Medical Informatics, Akdeniz University, Antalya, Turkey

[3]Department of Hospital Medicine, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Republic of Korea

[4]Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Republic of Korea

[5]Institute for Innovation in Digital Healthcare, Severance Hospital, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Dukyong Yoon, MD, PhD
Department of Biomedical Systems Informatics
Yonsei University College of Medicine
363, Dongbaekjukjeon-daero
Yongin, 16995
Republic of Korea
Phone: 82 3151898450
Email: dukyong.yoon@yonsei.ac.kr

Abstract

Background: The early prediction of antibiotic resistance in patients with a urinary tract infection (UTI) is important to guide appropriate antibiotic therapy selection.

Objective: In this study, we aimed to predict antibiotic resistance in patients with a UTI. Additionally, we aimed to interpret the machine learning models we developed.

Methods: The electronic medical records of patients who were admitted to Yongin Severance Hospital, South Korea were used. A total of 71 features extracted from patients' admission, diagnosis, prescription, and microbiology records were used for classification. UTI pathogens were classified as either sensitive or resistant to cephalosporin, piperacillin-tazobactam (TZP), carbapenem, trimethoprim-sulfamethoxazole (TMP-SMX), and fluoroquinolone. To analyze how each variable contributed to the machine learning model's predictions of antibiotic resistance, we used the Shapley Additive Explanations method. Finally, a prototype machine learning–based clinical decision support system was proposed to provide clinicians the resistance probabilities for each antibiotic.

Results: The data set included 3535, 737, 708, 1582, and 1365 samples for cephalosporin, TZP, TMP-SMX, fluoroquinolone, and carbapenem resistance prediction models, respectively. The area under the receiver operating characteristic curve values of the random forest models were 0.777 (95% CI 0.775-0.779), 0.864 (95% CI 0.862-0.867), 0.877 (95% CI 0.874-0.880), 0.881 (95% CI 0.879-0.882), and 0.884 (95% CI 0.884-0.885) in the training set and 0.638 (95% CI 0.635-0.642), 0.630 (95% CI 0.626-0.634), 0.665 (95% CI 0.659-0.671), 0.670 (95% CI 0.666-0.673), and 0.721 (95% CI 0.718-0.724) in the test set for predicting resistance to cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone, respectively. The number of previous visits, first culture after admission, chronic lower respiratory diseases, administration of drugs before infection, and exposure time to these drugs were found to be important variables for predicting antibiotic resistance.

Conclusions: The study results demonstrated the potential of machine learning to predict antibiotic resistance in patients with a UTI. Machine learning can assist clinicians in making decisions regarding the selection of appropriate antibiotic therapy in patients with a UTI.

(JMIR Med Inform 2024;12:e51326) doi:10.2196/51326

KEYWORDS

antibiotic resistance; machine learning; urinary tract infections; UTI; decision support

Introduction

Urinary tract infection (UTI) refers to an infection that occurs in any part of the urinary system, including the kidneys, ureters, urinary bladder, urethra, and other auxiliary structures [1,2]. Globally, UTIs are the most prevalent type of infectious disease, with around 150-250 million cases occurring each year [3]. Considerable morbidity and mortality result from these infections [4]. Typically, the most effective treatment for UTIs is the administration of antibiotics [3]. However, inappropriate use of antibiotics can permanently affect the normal microbiota of the urinary tract system and lead to antibiotic resistance [5].

The antibiotic susceptibility test is commonly used to identify antibiotic resistance, but it takes 24-48 hours to obtain test results [6,7]. However, in the clinical workflow, clinicians need to identify antibiotic resistance quickly to provide effective treatment for patients with UTIs. For this reason, early prediction of antibiotic resistance in patients with UTIs is important to guide the selection of appropriate antibiotic therapy. Machine learning can be used to develop prediction models and clinical decision support systems (CDSSs) to identify antibiotic resistance and support the selection of appropriate antibiotic therapy for patients with a UTI.

Several efforts have been made to predict antibiotic resistance in patients with UTIs using data from patients' electronic medical records (EMRs), including demographics, prescriptions, comorbidities, procedures, and laboratory tests. These investigations have yielded promising results. Some of these studies were limited to specific patient groups, including patients with uncomplicated UTIs [8] and patients treated in the emergency department [9]. In other studies, researchers worked with heterogeneous data that were not limited to a specific patient group [10-12]. However, prior studies that analyzed heterogeneous data did not address the interpretation of machine learning models. The black-box nature of machine learning is a limiting factor not only in its use for antibiotic resistance prediction but also in its wider clinical use [13,14]. Thus, interpreting the results obtained by the machine learning model is crucial in increasing users' trust in the machine learning model [15,16]. Furthermore, these studies did not address the development of the CDSS with the prediction models they built.

In this study, we aimed to predict antibiotic resistance in patients with a UTI. Heterogeneous data that were not limited to a specific patient group were used. UTI pathogens were classified as either sensitive or resistant to 5 commonly used antibiotics in UTI treatment: cephalosporin, piperacillin-tazobactam (TZP), carbapenem, trimethoprim-sulfamethoxazole (TMP-SMX), and fluoroquinolone. In addition, our objective was to understand and explain the inner workings of the machine learning models we developed. Eventually, a prototype CDSS was developed to provide clinicians the resistance probabilities for each antibiotic.

Methods

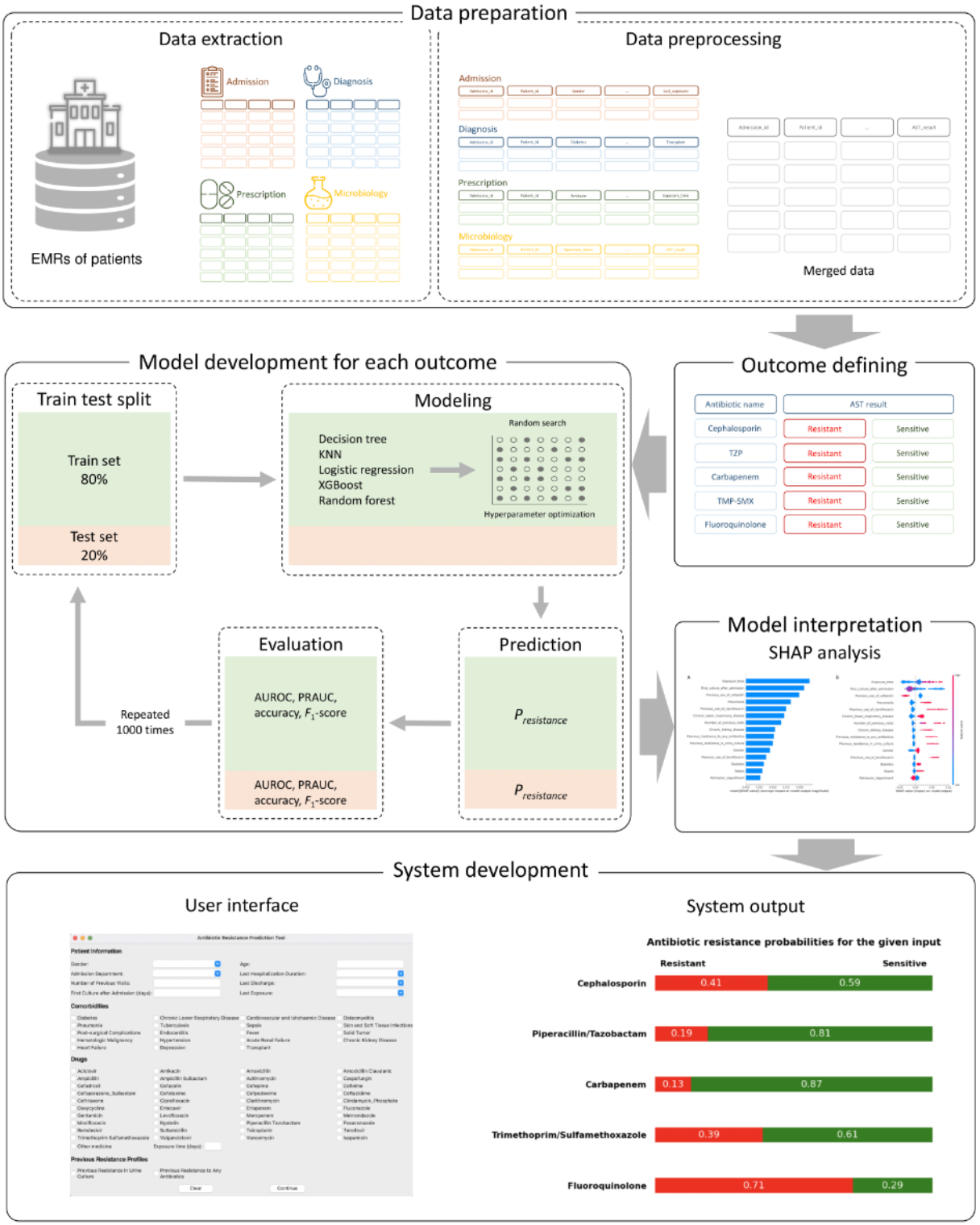
Ethical Considerations

Ethics approval for the study was obtained from the institutional review board of Yonsei University Severance Hospital on June 6, 2022 (approval 9-2023-0095). The informed consent was not required due to the retrospective nature of the study.

Data Set Description and Study Design

In this study, we used the EMRs of patients who were admitted to Yongin Severance Hospital, South Korea, between October 2012 and October 2022. To build the prediction models, admission, diagnosis, prescription, and microbiology records were extracted. The summary of the research process is presented in Figure 1.

Figure 1. Summary of the research process. AST: antibiotic susceptibility test; AUROC: area under the receiver operating characteristic curve; EMR: electronic medical record; KNN: k-nearest neighbor; PRAUC: precision-recall area under the curve; SHAP: Shapley Additive Explanations; TMP-SMX: trimethoprim-sulfamethoxazole; TZP: piperacillin-tazobactam; XGBoost: Extreme Gradient Boosting.



Data Preprocessing

The microbiology table contained 143,114 urine cultures collected from 6011 patients during 7719 admissions. Since positive samples typically indicate the presence of bacteriuria, and urine culture samples were typically collected from patients with UTI symptoms, we considered these to be indicative of a UTI [10]. The resistance profiles were evaluated based on the Clinical and Laboratory Standards Institute guidelines, where intermediate-level resistance was considered sensitive. To assess the resistance of UTI pathogens to antibiotic classes, antibiotics were grouped as cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone. The antibiotics included in each antibiotic class are presented in Multimedia Appendix 1. The patients' demographic information was extracted from the admission

table, their comorbidities were extracted from the diagnosis table, and their drug use information was extracted from the prescription table. For all input variables, the time of the first culture test was considered as the end point, and only data collected before the first culture test were used. After preprocessing and variable extraction from the raw data, the tables were combined using the admission number as the primary key. Missing data were excluded from the study. Patients aged 19 years and older and 100 years and younger at admission were included in the study, and numerical variables were standardized. A total of 71 features were used to classify UTI pathogens as either sensitive or resistant to each antibiotic. The predictors for the prediction models were selected by considering related works and using clinical judgment. Additionally, the threshold values for binarization were selected according to the literature [17] and the expert assessment of a specialist in infectious diseases. Detailed information about the predictors can be found in Multimedia Appendix 2.

Machine Learning Model Development

We used a repeated train test split approach for modeling. The data sets were split into training and test sets using an 80:20 ratio, and the training sets were used for the development of the machine learning models. When splitting the data into training and test sets, data points from the same patient and admission were exclusively included in either the training or test data set to prevent potential data leakage and ensure the models were evaluated on previously unseen data. At each iteration, we created different training and test data sets by changing the random seed. Decision tree, k-nearest neighbor, logistic regression, Extreme Gradient Boosting, and random forest were used for modeling. The hyperparameters of the machine learning models were optimized by using the random search hyperparameter optimization method with 10-fold cross-validation on the training data set. We stored the performance of the prediction models at each iteration, and the mean of performance metrics was calculated. The procedure of splitting the data, optimizing hyperparameters, modeling, and evaluation was iteratively repeated 1000 times to classify UTI pathogens as either sensitive or resistant to cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone. The machine learning models were built using Python (version 3.10.4; Python Software Foundation).

Machine Learning Model Interpretation

To analyze the contribution of the variables to the machine learning models in predicting antibiotic resistance, we used the Shapley Additive Explanations (SHAP) method. The SHAP values of the random forest models that showed superior performance compared to other machine learning methods were evaluated. The random forest model with the highest area under the receiver operating characteristic curve (AUROC) on the test set across all iterations for each antibiotic was used for SHAP analysis. Python (version 3.10.4; Python Software Foundation) was used for SHAP analysis.

CDSS Development

To develop the CDSS prototype, the random forest model with the highest AUROC on the test set across all iterations for each antibiotic was used. The CDSS prototype was developed using the *tkinter* package in Python (version 3.10.4; Python Software Foundation).

Evaluation

The performance of the machine learning model for predicting antibiotic resistance was evaluated on the training and test sets using the AUROC with 95% CIs, precision-recall area under the curve (PRAUC), accuracy, and F_1-score performance metrics. Herein, the AUROC value was considered the main evaluation metric. The definitions of the performance metrics we used are provided below.

- AUROC: The AUROC is a widely used metric that represents a classifier's ability to discriminate between positive instances and negative instances [18].
- PRAUC: PRAUC refers to the area under the precision-recall curve that plots precision as a function of recall for all the possible decision thresholds [19].
- Accuracy: Accuracy is the ratio of correctly classified samples to all samples.

- F_1-score: F_1-score is the harmonic mean of precision and recall metrics.

Python (version 3.10.4; Python Software Foundation) was used to evaluate the prediction models.

Results

Data Set Characteristics

The general characteristics of the data set used in this study are presented in Table 1. The data set included 3535, 737, 708, 1582, and 1365 samples for cephalosporin, TZP, TMP-SMX, fluoroquinolone, and carbapenem resistance prediction models, respectively. *Escherichia coli* was the most frequently isolated bacterial specimen across all antibiotics.

Table 1. General characteristics of the data set.

	Cephalosporin	TZP[a]	TMP-SMX[b]	Fluoroquinolone	Carbapenem
Samples, n	3535	737	708	1582	1365
Admissions, n	396	366	374	571	392
Patients, n	390	360	368	557	386
Resistance, n (%)	1492 (42.2)	169 (22.9)	281 (39.7)	1014 (64.1)	142 (10.4)
Age (years), mean (SD)	71.5 (14.4)	71.4 (14.4)	71.4 (14.4)	71.9 (14.4)	71.7 (14.3)
Female, n (%)	2597 (73.5)	523 (71)	507 (71.6)	1013 (64)	994 (72.8)
Most common bacteria (*Escherichia coli*), n (%)	1650 (46.7)	312 (42.3)	331 (46.7)	349 (22)	624 (45.7)
Second-most common bacteria (*Klebsiella pneumoniae*), n (%)	556 (15.7)	109 (14.8)	111 (15.7)	305 (19.3)[c]	220 (16.1)
Third-most common bacteria (*Pseudomonas aeruginosa*), n (%)	168 (4.7)	69 (9.4)	21 (3)[d]	180 (11.4)[e]	83 (6.1)

[a]TZP: piperacillin-tazobactam.

[b]TMP-SMX: trimethoprim-sulfamethoxazole.

[c]The isolated bacterial specimen is *Enterococcus faecium*.

[d]The isolated bacterial specimen is *Citrobacter freundii*.

[e]The isolated bacterial specimen is *Enterococcus faecalis*.

Model Performance

The performance analysis of the random forest models is presented in Table 2. The AUROC values were 0.777 (95% CI 0.775-0.779), 0.864 (95% CI 0.862-0.867), 0.877 (95% CI 0.874-0.880), 0.881 (95% CI 0.879-0.882), and 0.884 (95% CI 0.884-0.885) in the training set and 0.638 (95% CI 0.635-0.642), 0.630 (95% CI 0.626-0.634), 0.665 (95% CI 0.659-0.671), 0.670 (95% CI 0.666-0.673), and 0.721 (95% CI 0.718-0.724) in the test set for predicting resistance to cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone, respectively. The performance analysis of the other machine learning models is presented in Multimedia Appendices 3-6.

Table 2. Classification performances of the random forest models.

	Training set				Test set			
	AUROC[a] (95% CI)	PRAUC[b]	Accuracy	F_1-score	AUROC (95% CI)	PRAUC	Accuracy	F_1-score
Cephalosporin	0.777 (0.775-0.779)	0.725	0.715	0.676	0.638 (0.635-0.642)	0.547	0.603	0.556
TZP[c]	0.864 (0.862-0.867)	0.688	0.808	0.652	0.630 (0.626-0.634)	0.332	0.641	0.313
Carbapenem	0.877 (0.874-0.880)	0.539	0.822	0.493	0.665 (0.659-0.671)	0.222	0.725	0.220
TMP-SMX[d]	0.881 (0.879-0.882)	0.829	0.822	0.781	0.670 (0.666-0.673)	0.568	0.638	0.560
Fluoroquinolone	0.884 (0.884-0.885)	0.938	0.802	0.832	0.721 (0.718-0.724)	0.813	0.657	0.706

[a]AUROC: area under the receiver operating characteristic curve.

[b]PRAUC: precision-recall area under the curve.

[c]TZP: piperacillin-tazobactam.

[d]TMP-SMX: trimethoprim-sulfamethoxazole.

Important Features

The SHAP values of the 15 most important features in the random forest models are presented in Figure 2.

The SHAP feature importance bar plot (Figure 3A) and SHAP summary plot (Figure 3B) of the fluoroquinolone resistance prediction model are presented in Figure 3. The SHAP feature importance plot and SHAP summary plot of the other antibiotic prediction models are presented in Multimedia Appendices 7-10.

Figure 2. SHAP values of the 15 most important features in the prediction models. SHAP: Shapley Additive Explanations; TMP-SMX: trimethoprim-sulfamethoxazole; TZP: piperacillin-tazobactam.

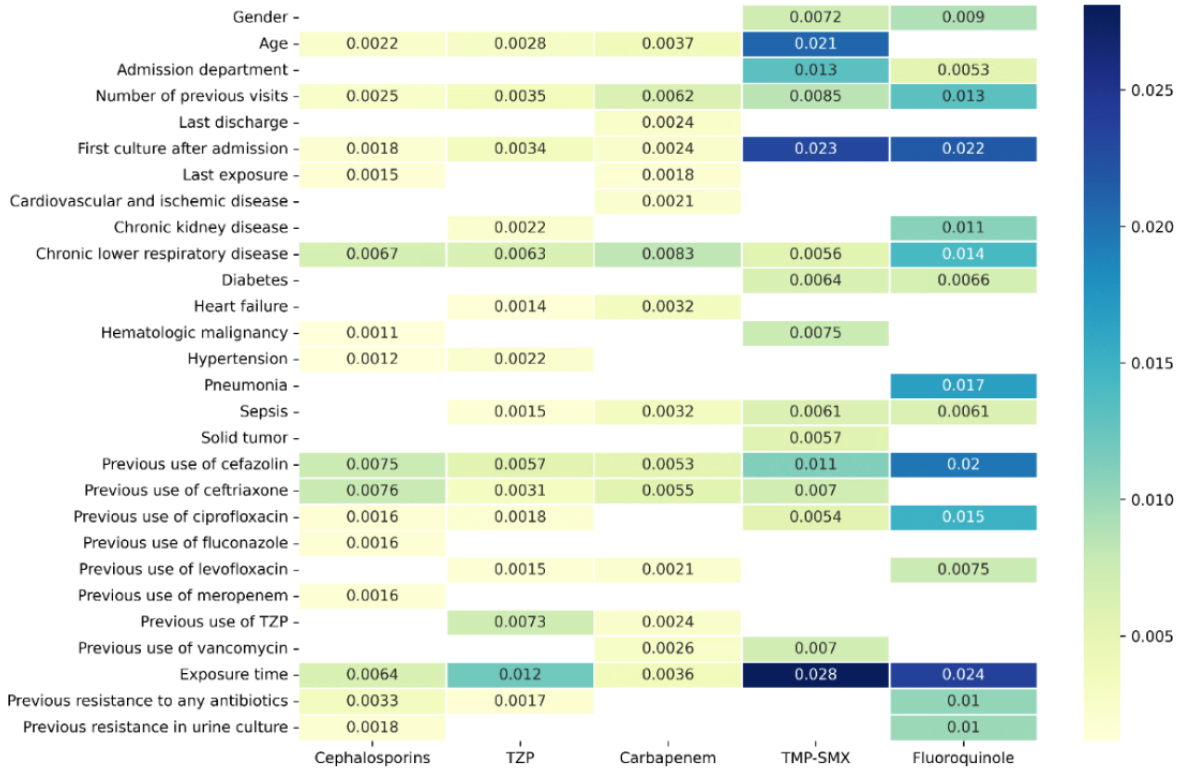
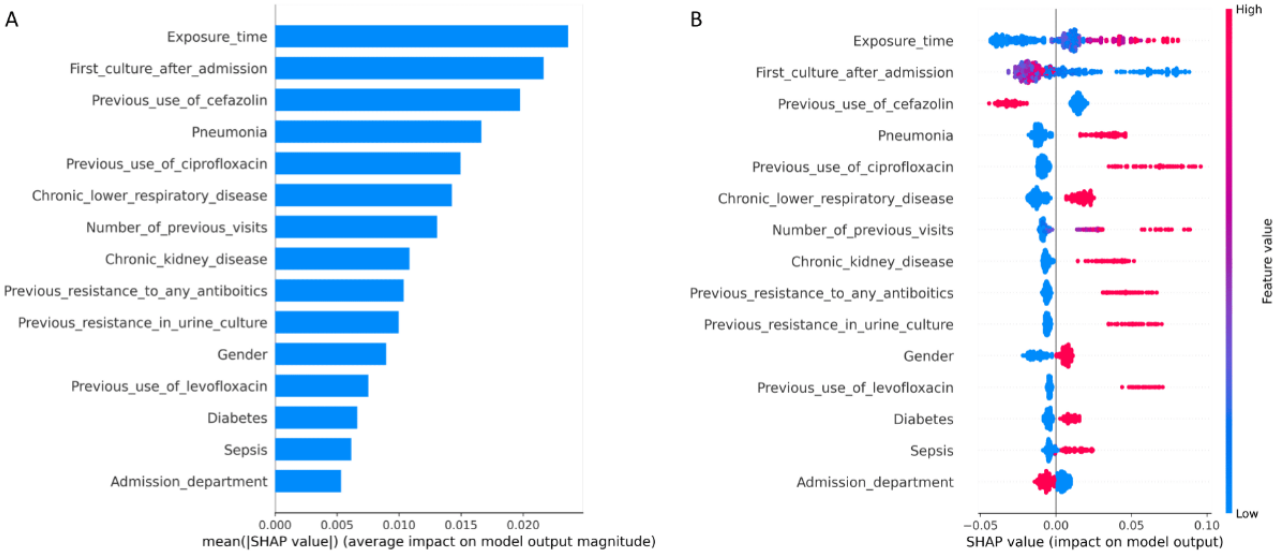


Figure 3. SHAP analysis results of fluoroquinolone resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations.



Clinical Decision Support System

The user interface of the CDSS is shown in Figure 4. The CDSS prototype obtains data from the user and produces antibiotic resistance probabilities for each antibiotic.

We presented the CDSS prototype on a scenario. In this case, a female aged 55 years was admitted to the hospital's outpatient department. The patient previously visited the hospital 3 times and was readmitted to the hospital within 30 days of her last 3-day stay. The duration between the patient's admission to the hospital and the first culture was 1 day. The patient was previously diagnosed with diabetes and chronic lower respiratory disease. Additionally, the patient had a history of cefazolin use in the last 30 days and resistance in urine culture.

The system output for the given scenario is shown in Figure 5. The system produced resistance probabilities for each antibiotic. For the given scenario, the system produced a 71% probability of fluoroquinolone resistance, a 41% probability of cephalosporin resistance, a 39% probability of TMP-SMX resistance, a 19% probability of TZP resistance, and a 13% probability of carbapenem resistance.

Figure 4. The user interface of the clinical decision support system.

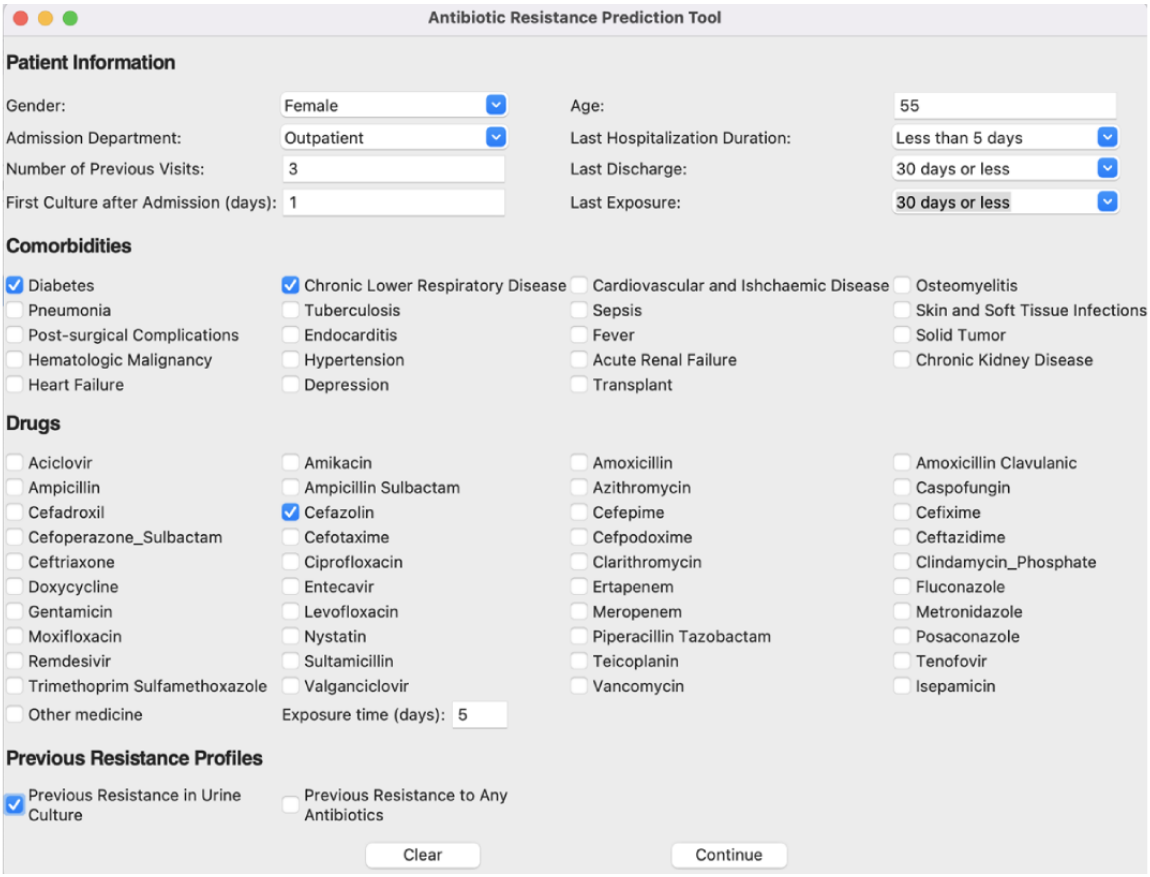
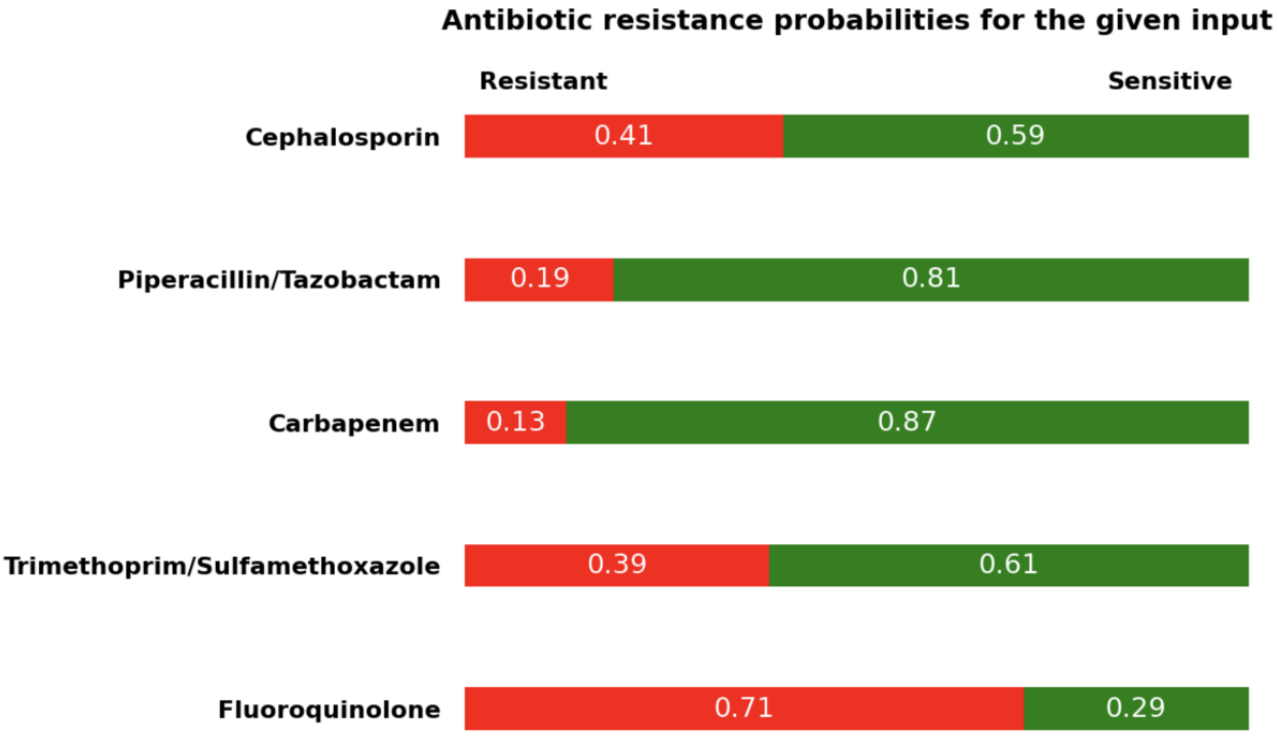


Figure 5. The screenshot of system output for the given data.



Discussion

Principal Findings

In this study, our main objective was to predict cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone resistance in patients with UTI and develop a CDSS with the machine learning models we built. Moreover, we identified the most important features for predicting antibiotic resistance in patients with UTI using SHAP analysis.

Our prediction models achieved AUROCs of 0.777 (95% CI 0.775-0.779), 0.864 (95% CI 0.862-0.867), 0.877 (95% CI 0.874-0.880), 0.881 (95% CI 0.879-0.882), and 0.884 (95% CI 0.884-0.885) in the training set and 0.638 (95% CI 0.635-0.642), 0.630 (95% CI 0.626-0.634), 0.665 (95% CI 0.659-0.671), 0.670 (95% CI 0.666-0.673), and 0.721 (95% CI 0.718-0.724) in the test set for predicting resistance to cephalosporin, TZP, carbapenem, TMP-SMX, and fluoroquinolone, respectively. The fluoroquinolone resistance prediction model showed superior performance, as confirmed by its high AUROC values in both the training and test sets. On the other hand, the cephalosporin resistance prediction model showed poor performance, as confirmed by the low AUROC values in both training and test sets.

According to SHAP analysis, the contribution of the variables varied for each antibiotic; however, we found that the number of previous visits, first culture after admission, chronic lower respiratory diseases, administration of drugs before infection, and exposure time to these drugs were important predictors across all antibiotics. Factors such as the first culture after admission, exposure time, and the number of previous visits were found to affect resistance, which can be explained by the impact of health care–associated infections. Chronic lower respiratory and kidney diseases are also likely to be associated with frequent visits to health care facilities, although it is difficult to confirm the actual number of visits. However, this suggests that the characteristics of health care–seeking behavior in patients with specific underlying diseases may influence resistance [20]. Interestingly, the use of cefazolin had a negative impact on the development of resistance for all antibiotics. This is because cefazolin is one of the narrow-spectrum antibiotics used in less severe patients. Further research is needed to examine these results.

Comparison to Prior Work

Past efforts to predict antibiotic resistance in patients with UTIs have had promising results, with the lowest AUROC being 0.58 for predicting TMP-SMX resistance [12] and the highest AUROC being 0.83 for predicting ciprofloxacin resistance [9]. In comparison, our prediction models demonstrated comparable performance to these prior works. Some previous studies on predicting antibiotic resistance in patients with UTIs were limited to specific patient groups, including patients with uncomplicated UTIs [8] and patients treated in the emergency department [9]. We analyzed heterogeneous data that were not limited to a specific patient group or bacteria. This approach provides a more comprehensive insight into the prediction of antibiotic resistance in patients with UTIs. Similarly, Lewin-Epstein et al [21] analyzed heterogeneous data and were able to achieve AUROC values ranging from 0.73 to 0.79 for the prediction of ceftazidime, gentamicin, imipenem, ofloxacin, and TMP-SMX resistance. Their data contained multiple culture tests, which provided a more comprehensive approach to predicting antibiotic resistance. Although urine cultures can be used to infer colonized resistance in patients, further research is needed to extend culture results beyond urine.

Limitations

While this study provides insights into predicting antibiotic resistance in patients with UTIs, it has some limitations. First, this study is the lack of multidrug resistance classification. The data set we used in this study did not contain a sufficient amount of multidrug resistance outcomes to build a classification model for the prediction of multidrug resistance. Furthermore, our prediction models were developed using prescription records within the hospital setting. However, patients may have used antibiotics outside of the hospital setting during visits to other hospitals. The lack of information about past drug use could have negatively impacted the performance of our prediction models. To overcome this limitation, we intend to conduct further studies using data from the National Health Insurance Service of South Korea, which contain all past drug use information of the patients. Thus, we will have a more comprehensive data set. By using this approach, we may be able to develop more accurate machine learning models to predict antibiotic resistance and improve our ability to guide appropriate antibiotic therapy selection. Additionally, further development is required to address the limitations of prototype CDSS, including the integration of real-time patient data and validation in larger patient cohorts. Moreover, the prototype CDSS only gives the resistance risk probability to the user. However, a more comprehensive system that can provide decision support on the selection of appropriate therapy, dosage, and duration of treatment can be developed in further studies. Such a system has the potential to reduce the duration of treatment, number of antibiotics used, cost, mortality, and morbidity [22,23].

Conclusions

In conclusion, our study results demonstrated that prediction models to predict antibiotic resistance in patients with UTIs can be constructed using routinely collected EMR data alone, without requiring additional laboratory tests or specialized tests. Machine learning techniques can be used to develop systems that can guide clinicians in selecting appropriate antibiotic therapy. This has the potential to prevent the risk of inappropriate antibiotic administration, thereby reducing patients' risk of developing antibiotic resistance.

Acknowledgments

This study was supported by the National Institute for International Education of the Government of the Republic of Korea, The Scientific and Technological Research Council of Turkey (grant 2214-A), and a faculty research grant of Yonsei University College of Medicine (6-2022-0118).

Authors' Contributions

Nİ, AY, SYP, and DY contributed to the conceptualization of the study and to the funding acquisition. JK, JAR, and SYP were responsible for data curation. Nİ performed the formal analysis of the collected data and wrote the paper. Nİ, SYP, and DY contributed to the development of the study methodology. SYP and DY reviewed and edited the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The antibiotics included in each antibiotic class.
[DOCX File , 17 KB - medinform_v12i1e51326_app1.docx]

Multimedia Appendix 2
Description of input variables.
[DOCX File , 15 KB - medinform_v12i1e51326_app2.docx]

Multimedia Appendix 3
Classification performances of the decision tree models.
[DOCX File , 20 KB - medinform_v12i1e51326_app3.docx]

Multimedia Appendix 4
Classification performances of the k-nearest neighbor models.
[DOCX File , 20 KB - medinform_v12i1e51326_app4.docx]

Multimedia Appendix 5
Classification performances of the logistic regression models.
[DOCX File , 20 KB - medinform_v12i1e51326_app5.docx]

Multimedia Appendix 6
Classification performances of the Extreme Gradient Boosting models.
[DOCX File , 20 KB - medinform_v12i1e51326_app6.docx]

Multimedia Appendix 7
SHAP analysis results of cephalosporin resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations.
[PNG File , 121 KB - medinform_v12i1e51326_app7.png]

Multimedia Appendix 8
SHAP analysis results of TZP resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations; TZP: piperacillin-tazobactam.
[PNG File , 117 KB - medinform_v12i1e51326_app8.png]

Multimedia Appendix 9
SHAP analysis results of carbapenem resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations; TZP: piperacillin-tazobactam.
[PNG File , 122 KB - medinform_v12i1e51326_app9.png]

Multimedia Appendix 10
SHAP analysis results of TMP-SMX resistance prediction model. (A) The feature importance bar plot. (B) The SHAP summary dot plot. SHAP: Shapley Additive Explanations; TMP-SMX: trimethoprim-sulfamethoxazole.

[PNG File , 123 KB - medinform_v12i1e51326_app10.png]

References

1. Tan CW, Chlebicki MP. Urinary tract infections in adults. Singapore Med J 2016;57(9):485-490 [FREE Full text] [doi: 10.11622/smedj.2016153] [Medline: 27662890]
2. Belete MA, Saravanan M. A systematic review on drug resistant urinary tract infection among pregnant women in developing countries in Africa and Asia; 2005-2016. Infect Drug Resist 2020;13:1465-1477 [FREE Full text] [doi: 10.2147/IDR.S250654] [Medline: 32547115]
3. Santos M, Mariz M, Tiago I, Martins J, Alarico S, Ferreira P. A review on urinary tract infections diagnostic methods: laboratory-based and point-of-care approaches. J Pharm Biomed Anal 2022;219:114889 [FREE Full text] [doi: 10.1016/j.jpba.2022.114889] [Medline: 35724611]
4. Suskind AM, Saigal CS, Hanley JM, Lai J, Setodji CM, Clemens JQ, Urologic Diseases of America Project. Incidence and management of uncomplicated recurrent urinary tract infections in a national sample of women in the United States. Urology 2016;90:50-55 [FREE Full text] [doi: 10.1016/j.urology.2015.11.051] [Medline: 26825489]
5. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. Nat Rev Microbiol 2015;13(5):269-284 [FREE Full text] [doi: 10.1038/nrmicro3432] [Medline: 25853778]
6. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet 2019;20(6):356-370 [FREE Full text] [doi: 10.1038/s41576-019-0108-4] [Medline: 30886350]
7. Benkova M, Soukup O, Marek J. Antimicrobial susceptibility testing: currently used methods and devices and the near future in clinical practice. J Appl Microbiol 2020;129(4):806-822. [doi: 10.1111/jam.14704] [Medline: 32418295]
8. Kanjilal S, Oberst M, Boominathan S, Zhou H, Hooper DC, Sontag D. A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. Sci Transl Med 2020;12(568):eaay5067 [FREE Full text] [doi: 10.1126/scitranslmed.aay5067] [Medline: 33148625]
9. Lee HG, Seo Y, Kim JH, Han SB, Im JH, Jung CY, et al. Machine learning model for predicting ciprofloxacin resistance and presence of ESBL in patients with UTI in the ED. Sci Rep 2023;13(1):3282 [FREE Full text] [doi: 10.1038/s41598-023-30290-y] [Medline: 36841917]
10. Yelin I, Snitser O, Novich G, Katz R, Tal O, Parizade M, et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. Nat Med 2019;25(7):1143-1152 [FREE Full text] [doi: 10.1038/s41591-019-0503-6] [Medline: 31273328]
11. Hebert C, Gao Y, Rahman P, Dewart C, Lustberg M, Pancholi P, et al. Prediction of antibiotic susceptibility for urinary tract infection in a hospital setting. Antimicrob Agents Chemother 2020;64(7):e02236-19 [FREE Full text] [doi: 10.1128/AAC.02236-19] [Medline: 32312778]
12. Rich SN, Jun I, Bian J, Boucher C, Cherabuddi K, Morris JG, et al. Development of a prediction model for antibiotic-resistant urinary tract infections using integrated electronic health records from multiple clinics in North-Central Florida. Infect Dis Ther 2022;11(5):1869-1882 [FREE Full text] [doi: 10.1007/s40121-022-00677-x] [Medline: 35908268]
13. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886. [doi: 10.1136/bmj.l886] [Medline: 30862612]
14. Macesic N, Polubriaginof F, Tatonetti NP. Machine learning: novel bioinformatics approaches for combating antimicrobial resistance. Curr Opin Infect Dis 2017;30(6):511-517. [doi: 10.1097/QCO.0000000000000406] [Medline: 28914640]
15. Tucci V, Saary J, Doyle TE. Factors influencing trust in medical artificial intelligence for healthcare professionals: a narrative review. J Med Artif Intell 2022;5:4-4 [FREE Full text] [doi: 10.21037/jmai-21-25]
16. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019;19(1):146 [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]
17. Tacconelli E. New strategies to identify patients harbouring antibiotic-resistant bacteria at hospital admission. Clin Microbiol Infect 2006;12(2):102-109 [FREE Full text] [doi: 10.1111/j.1469-0691.2005.01326.x] [Medline: 16441446]
18. Janssens ACJW, Martens FK. Reflection on modern methods: revisiting the area under the ROC curve. Int J Epidemiol 2020;49(4):1397-1403 [FREE Full text] [doi: 10.1093/ije/dyz274] [Medline: 31967640]
19. Cook J, Ramadas V. When to consult precision-recall curves. Stata J 2020;20(1):131-148 [FREE Full text] [doi: 10.1177/1536867x20909693]
20. Park H, Son MJ, Jung DW, Lee H, Lee JY. National trends in hospitalization for ambulatory care sensitive conditions among Korean adults between 2008 and 2019. Yonsei Med J 2022;63(10):948-955 [FREE Full text] [doi: 10.3349/ymj.2022.0110] [Medline: 36168248]
21. Lewin-Epstein O, Baruch S, Hadany L, Stein GY, Obolski U. Predicting antibiotic resistance in hospitalized patients by applying machine learning to electronic medical records. Clin Infect Dis 2021;72(11):e848-e855 [FREE Full text] [doi: 10.1093/cid/ciaa1576] [Medline: 33070171]
22. Curtis CE, Al Bahar F, Marriott JF. The effectiveness of computerised decision support on antibiotic use in hospitals: a systematic review. PLoS One 2017;12(8):e0183062 [FREE Full text] [doi: 10.1371/journal.pone.0183062] [Medline: 28837665]

23. Laka M, Milazzo A, Merlin T. Can evidence-based decision support tools transform antibiotic management? A systematic review and meta-analyses. J Antimicrob Chemother 2020;75(5):1099-1111 [FREE Full text] [doi: 10.1093/jac/dkz543] [Medline: 31960021]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CDSS: clinical decision support system
EMR: electronic medical record
PRAUC: precision-recall area under the curve
SHAP: Shapley Additive Explanations
TMP-SMX: trimethoprim-sulfamethoxazole
TZP: piperacillin-tazobactam
UTI: urinary tract infection

Edited by A Benis; submitted 27.07.23; peer-reviewed by MO Khursheed, YJ Tseng; comments to author 25.08.23; revised version received 17.11.23; accepted 08.01.24; published 29.02.24.

Please cite as:
İlhanlı N, Park SY, Kim J, Ryu JA, Yardımcı A, Yoon D
Prediction of Antibiotic Resistance in Patients With a Urinary Tract Infection: Algorithm Development and Validation
JMIR Med Inform 2024;12:e51326
URL: https://medinform.jmir.org/2024/1/e51326
doi:10.2196/51326
PMID:38421718

Original Paper

Forecasting Hospital Room and Ward Occupancy Using Static and Dynamic Information Concurrently: Retrospective Single-Center Cohort Study

Hyeram Seo[1], BS; Imjin Ahn[2], MS; Hansle Gwon[2], MS; Heejun Kang[3], MS; Yunha Kim[2], MS; Heejung Choi[2], MS; Minkyoung Kim[1], BS; Jiye Han[1], BS; Gaeun Kee[2], MS; Seohyun Park[2], BS; Soyoung Ko[2], BS; HyoJe Jung[2], BS; Byeolhee Kim[2], BS; Jungsik Oh[4], BS; Tae Joon Jun[5*], PhD; Young-Hak Kim[6*], MD, PhD

[1]Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center & University of Ulsan College of Medicine, Seoul, Republic of Korea

[2]Department of Information Medicine, Asan Medical Center, Seoul, Republic of Korea

[3]Division of Cardiology, Asan Medical Center, Seoul, Republic of Korea

[4]Department of Digital Innovation, Asan Medical Center, Seoul, Republic of Korea

[5]Big Data Research Center, Asan Institute for Life Sciences, Asan Medical Center, Seoul, Republic of Korea

[6]Division of Cardiology, Department of Information Medicine, Asan Medical Center & University of Ulsan College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Young-Hak Kim, MD, PhD
Division of Cardiology
Department of Information Medicine
Asan Medical Center & University of Ulsan College of Medicine
88, Olympic-ro 43-gil
Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 0955
Email: mdyhkim@amc.seoul.kr

Abstract

Background: Predicting the bed occupancy rate (BOR) is essential for efficient hospital resource management, long-term budget planning, and patient care planning. Although macro-level BOR prediction for the entire hospital is crucial, predicting occupancy at a detailed level, such as specific wards and rooms, is more practical and useful for hospital scheduling.

Objective: The aim of this study was to develop a web-based support tool that allows hospital administrators to grasp the BOR for each ward and room according to different time periods.

Methods: We trained time-series models based on long short-term memory (LSTM) using individual bed data aggregated hourly each day to predict the BOR for each ward and room in the hospital. Ward training involved 2 models with 7- and 30-day time windows, and room training involved models with 3- and 7-day time windows for shorter-term planning. To further improve prediction performance, we added 2 models trained by concatenating dynamic data with static data representing room-specific details.

Results: We confirmed the results of a total of 12 models using bidirectional long short-term memory (Bi-LSTM) and LSTM, and the model based on Bi-LSTM showed better performance. The ward-level prediction model had a mean absolute error (MAE) of 0.067, mean square error (MSE) of 0.009, root mean square error (RMSE) of 0.094, and R^2 score of 0.544. Among the room-level prediction models, the model that combined static data exhibited superior performance, with a MAE of 0.129, MSE of 0.050, RMSE of 0.227, and R^2 score of 0.600. Model results can be displayed on an electronic dashboard for easy access via the web.

Conclusions: We have proposed predictive BOR models for individual wards and rooms that demonstrate high performance. The results can be visualized through a web-based dashboard, aiding hospital administrators in bed operation planning. This contributes to resource optimization and the reduction of hospital resource use.

(JMIR Med Inform 2024;12:e53400) doi:10.2196/53400

KEYWORDS

hospital bed occupancy; electronic medical records; time series forecasting; short-term memory; combining static and dynamic variables

Introduction

Background

The global health care market continues to grow, but the burden of health care costs on governments and individuals is reaching its limits. Consequently, there is increasing interest in the efficient use of limited resources in health care systems, and hospitals must develop approaches to maximize medical effectiveness within budgetary constraints [1,2]. One approach to this is optimizing the use of medical resources. Medical resources can be broadly categorized into 3 categories: human resources, physical capital, and consumables. The appropriate and optimized use of these resources is critical for improving health care quality and providing care to a larger number of patients [3,4].

Among the 3 medical resources, hospital beds are considered one of the physical capitals provided by hospitals to patients. These beds are allocated for various purposes, such as rest, hospitalization, postsurgical recovery, etc. They constitute one of the factors that can directly influence the patient's internal satisfaction within the hospital. However, owing to limited space, hospitals often have a restricted number of beds. Moreover, the number and functionality of beds are often fixed owing to budgetary or environmental constraints, making it difficult to make changes. Nonetheless, if hospital administrators can evaluate bed occupancy rates (BORs) according to different time periods, they can predict the need for health care professionals and resources. On the basis of this information, hospitals can plan resources efficiently, reduce operational costs, and achieve economic objectives [5]. In addition, excessive BORs can exert a negative effect on the health of staff members and increase the possibility of exposure to infection risks. Hence, emphasizing only maintaining a high BOR may not necessarily lead to favorable outcomes for the hospital [6,7]. Considering these reasons, BOR prediction plays a vital role in hospitals and is recognized as a broadly understood necessity for resource optimization in the competitive medical field.

In the medical field, optimizing resources is crucial in the face of limited bed capacity and intense competition. Therefore, bed planning is a vital consideration aimed at minimizing hospital costs [8]. To achieve this, hospitals need to plan staffing and vacations weeks or months in advance [9]. The use of machine learning (ML) technology for BOR prediction is necessary to address fluctuations in patient numbers due to seasonal variations or infectious diseases, ensuring continuous hospital operations. In the Netherlands, hospitals have already implemented ML-based BOR prediction [10], and Johns Hopkins Hospital uses various metrics to effectively manage bed capacity for optimization. Predicting BORs based on quantitative data contributes to validating the clinical quality and cost-effectiveness of treatments. This, in turn, enhances overall accountability throughout the wards and contributes to improving hospital efficiency [11].

Prior Work

Hospital BOR prediction has been investigated using various approaches recently. From studies predicting bed demand using mathematical statistics or regression equation models based on given data [12-15], the focus has shifted toward modeling approaches using time-series analysis. This approach observes recorded data over time to predict future values.

A previous study has taken an innovative approach using time-series analysis alongside the commonly used regression analysis for bed demand prediction, and the study demonstrated that using time-series prediction for bed occupancy yielded higher performance results than using a simple trend fitting approach [16]. Another study used the autoregressive integrated moving average (ARIMA) model for univariate data and a time-series model for multivariate data to predict BORs [17]. With the advancement of deep learning (DL) models that possess strong long-term memory capabilities, such as recurrent neural network (RNN) and long short-term memory (LSTM), there has been an increase in studies applying these models to time-series data for prediction purposes. For instance, in the study by Kutafina et al [9], hospital BORs were predicted based on dates and public holiday data from government agencies and schools, without involving the personal information of patients. The study used a nonlinear autoregressive exogenous model to predict a short-term period of 60 days, with an aim to contribute to the planning of hospital staff. The model demonstrated good performance, with an average mean absolute percentage error of 6.24%. In emergency situations, such as the recent global COVID-19 pandemic, the sudden influx of infected patients can disrupt the hospitalization plans for patients with pre-existing conditions [18]. Studies have been conducted using DL architectures to design models for predicting the BOR of patients with COVID-19 on a country-by-country basis. Some studies incorporated additional inputs, such as vaccination rate and median age, to train the models [19]. Studies have also been conducted to focus on the short-term prediction of BORs during the COVID-19 period [20,21]. Prior studies are summarized in Table 1.

Although previous research has contributed to BOR prediction and operational planning at the hospital level, more detailed and systematic predictions are necessary for practical application in real-world operations. To address this issue, studies have developed their own computer simulation hospital systems to not only predict bed occupancy but also execute scheduling for admissions and surgeries to enhance resource utilization [22-24]. Nevertheless, existing studies have the limitation of focusing solely on the overall BOR of the hospital. As an advancement to these studies, we aim to propose a strategy for predicting the BOR at the level of each ward and room using various variables

in a time-series manner. Interestingly, to our knowledge, this is the first study to apply DL to predict ward- and room-specific occupancy rates using time-series analysis.

Table 1. Summary of prior studies.

Study	Year	Data set	Method	Prediction target
Mackay and Lee [12]	2007	Deidentified data, the date and time of patient admission and discharge between 1998 and 2000	Comparison of 2 compartment models through cross-validation	Entire hospital bed occupancy (annual average)
Littig and Isken [13]	2007	Historical and real-time data warehouse and hospital information systems (emergency department, financial, surgical scheduling, and inpatient tracking systems)	Computerized model of MLR[a] and LR[b]	Entire hospital short-term occupancy (24 h or 72 h) based on LOS[c]
Kumar and Mo [14]	2010	Bed management between June 1, 2006, and June 1, 2007; Information: (1) In each class based on length of stay and admission data; (2) Historical previous year's same week admission data; (3) Relationship between identified variables to aid bed managers	The 3 methods are: (1) Poisson bed occupancy model; (2) Simulation model; and (3) Regression model	The 3 prediction targets are: (1) Estimation of bed occupancy and optimal bed requirements in each class; (2) Bed occupancy levels for every class for the following week; and (3) Weekly average number of occupied beds
Seematter-Bagnoud et al [15]	2015	Inpatient stay data in 2010 (acute somatic care inpatients and outpatients)	Three models of hypothesis-based statistical forecasting of future trends	The 3 targets are: (1) Number of hospital stays; (2) Hospital inpatient days; and (3) Beds for medical stay
Farmer and Emami [16]	1990	Inpatient stay data for general surgery in the age group of 15-44 years between 1969 and 1982	The 2 methods are: (1) Forecasting from a structural model and (2) The time-series or Box-Jenkins method	Entire hospital short-term daily bed requirements
Kim et al [17]	2014	Data warehouse between January 2009 and June 2012	The 2 methods are: (1) The ARIMA[d] model for univariate data and (2) The time-series model for multivariate data	Entire hospital bed occupancy (1 day and 1 week)
Kutafina et al [9]	2019	Inpatient stay data between October 14, 2002, and December 31, 2015 (patient identifier, time of admission, discharge, and name of the clinic the patient was admitted to; no personal information on the patients or staff was provided)	NARX[e] model, a type of RNN[f]	Entire hospital mid-term bed occupancy (60 days, bed pool in units of 30 beds)
Bouhamed et al [19]	2022	COVID-19 hospital occupancy data in 15 countries between December 2021 and early January 2022	The 3 models are: LSTM[g], GRU[h], and SRNN[i]. Incorporate vaccination percentage and median age of the population to improve performance	Entire hospital bed occupancy
Bekker et al [20]	2021	Historical data publicly available until mid-October 2020	The 2 methods are: (1) Using linear programming to predict admissions and (2) Fitting the remaining LOS and using results from the queuing theory to predict occupancy	The 2 targets are: (1) Patient admission and (2) Entire hospital short-term bed occupancy
Farcomeni et al [21]	2021	Patients admitted to the intensive care unit between January and June 2020	The 2 methods are: (1) Generalized linear mixed regression model and (2) Area-specific nonstationary integer autoregressive methodology	Entire hospital short-term intensive care bed occupancy

[a]MLR: multinomial logistic regression.

[b]LR: linear regression.

[c]LOS: length of stay.

[d]ARIMA: autoregressive integrated moving average.

[e]NARX: nonlinear autoregressive exogenous.

[f]RNN: recurrent neural network.

[g]LSTM: long short-term memory.

[h]GRU: grid recurrent unit.

[i]SRNN: simple recurrent neural network.

Goal of This Study

The aim of this study was to predict the BORs of hospital wards and rooms using time-series data from individual beds. Although overall bed occupancy prediction is useful for macro-level resource management in hospitals, resource allocation based on the prediction of occupancy rates for each ward and room is required for specific hospital scheduling and practicality. Through this approach, we aim to contribute to the efficient operational cost optimization of the hospital and ensure the availability of resources required for patient care.

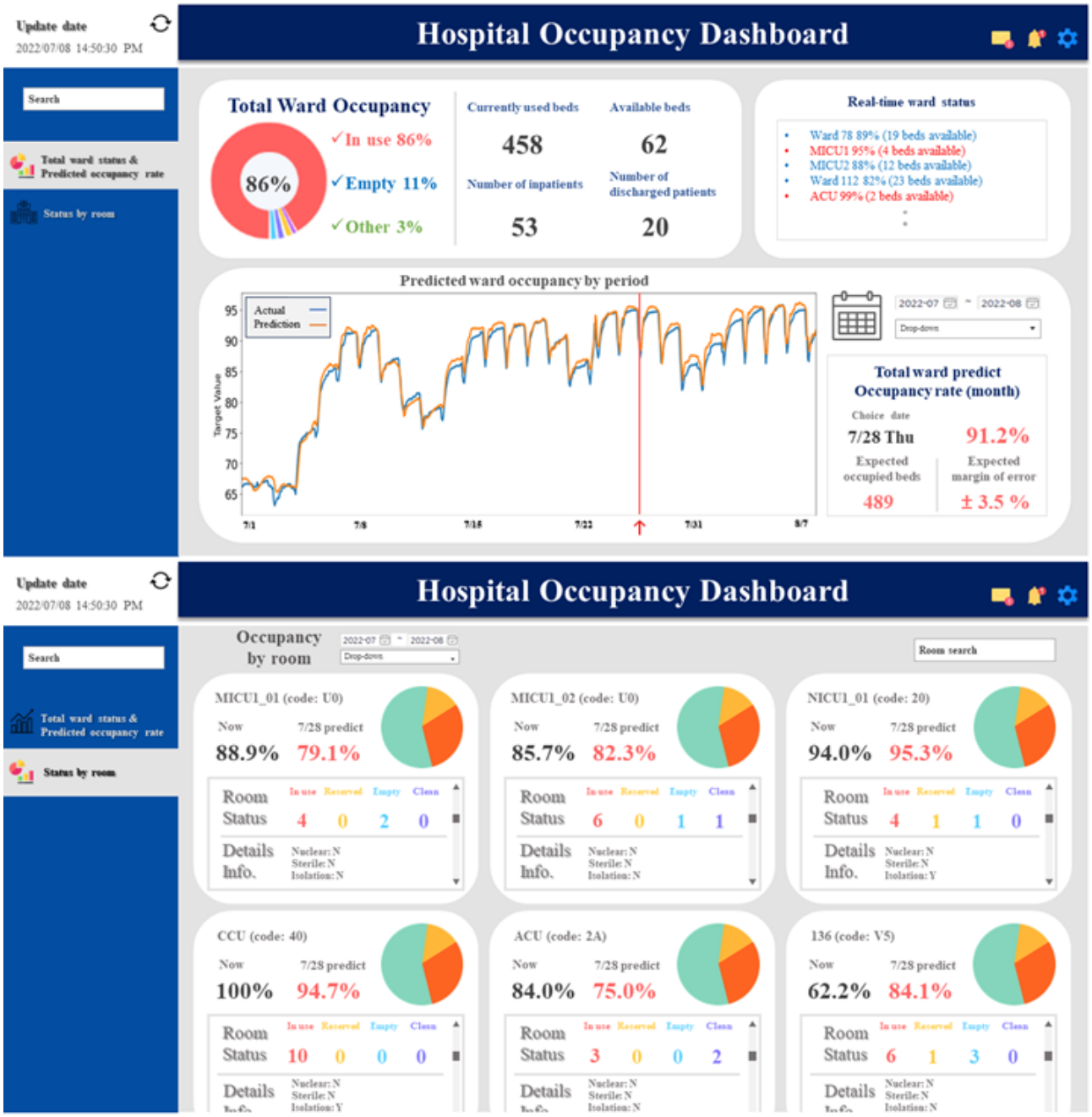
We have developed time-series prediction models based on deep neural network (DNN), among which 1 model combines data representing room-specific features (static data) with dynamic data to enhance the prediction performance for room bed occupancy rates (RBORs). Based on bidirectional long short-term memory (Bi-LSTM), the RBOR prediction model demonstrates a lower mean absolute error (MAE) of 0.049, a mean square error (MSE) of 0.042, a root mean square error (RMSE) of 0.007, and a higher R^2 score of 0.291, indicating the highest performance among all RBOR models.

We developed 6 types of BOR prediction models, of which 2 types were used for predicting ward bed occupancy rates (WBORs), and the other 4 types focused on predicting RBORs. These models use LSTM and Bi-LSTM architectures with strong long-term memory capabilities as their basic structure. We created 6 models for each architecture, resulting in a total of 12 models. The WBOR models were used for predicting weekly and monthly occupancy rates, serving long-term hospital administrative planning purposes. Conversely, the RBOR models were designed for immediate and rapid occupancy planning and were trained with 3- and 7-day intervals. Each RBOR model was enhanced by combining static data, which represent room-specific features, to generate more sophisticated prediction models.

Figure 1 shows the potential application of our model as a form of web software in a hospital setting. Through an online dashboard, it can provide timely information regarding bed availability, enabling intelligent management of patient movements related to admission and discharge. It facilitates shared responsibilities within the hospital and simplifies future resource planning [25].

In the Introduction section, we explored the importance of this research and investigated relevant previous studies, providing a general overview of the direction of our research. In the Methods section, we provide descriptions of the data set used and the structure of the DNN algorithm used, and explain the model architecture and performance. In the Results section, we present the performance and outcomes of this study. Finally, in the Discussion section, we discuss the contributions, limitations, and potential avenues for improvement of the research.

Figure 1. Virtual dashboard of the status and forecast of the ward bed occupancy rate (WBOR) and room bed occupancy rate (RBOR). The first screen presents the overall bed occupancy rate of the hospital, along with the number of beds in use and available. Moreover, a predictive graph displays the anticipated WBOR for selected dates. The second screen presents the WBOR for individual beds, indicating their statuses, such as "in use," "reserved," "empty," and "cleaning." Detailed information about each room is also displayed.



Methods

Overview

We intended to predict the BORs of individual hospital wards and rooms based on the information accumulated in individual bed–level data on an hourly basis, aggregated on a daily basis. For this purpose, we developed 12 time-series models. As the base models, we applied LSTM and Bi-LSTM, which are suitable for sequence data. These models address the limitation of long-term memory loss in traditional RNNs and were chosen because of their suitability for training bed data represented as sequence data.

Based on the model architecture, there were 2 WBOR prediction model types, which were trained at 7- and 30-day intervals to predict the occupancy rate for the next day. Moreover, there were 2 RBOR prediction model types, similar to the ward models, which were trained at 3- and 7-day intervals. Furthermore, as another approach, each RBOR prediction model was augmented with static data, and 2 DL algorithms were proposed for the final comparison of their performances in predicting RBORs.

Ethical Considerations

The study was approved by the Asan Medical Center (AMC) Institutional Review Board (IRB 2021-0321) and was conducted in accordance with the 2008 Declaration of Helsinki.

Materials

Study Setting

This was a retrospective single-center cohort study. Data were collected from AMC, with information on the occupancy status of each bed recorded at hourly intervals between May 27, 2020, and November 21, 2022. The data set comprised a total of 54,632,684 records. This study used ethically preapproved data. Deidentified data used in the study were extracted from ABLE, the AMC clinical research data warehouse.

A total of 57 wards, encompassing specialized wards; 1411 rooms, including private and shared rooms; and 4990 beds were included in this study. Wards and rooms with specific characteristics, such as intensive care unit, newborn room, and nuclear medicine treatment room, were excluded from the analysis as their occupancy prediction using simple and general variables did not align with the direction of this study.

Supporting Data

Supporting data for public holidays were added in our data set. We considered that holidays have both a recurring pattern with specific dates each year and a distinctive characteristic of being nonworking days, which could affect occupancy rates. Based on Korean public holidays, which include Chuseok, Hangeul Proclamation Day, Children's Day, National Liberation Day, Memorial Day, Buddha's Birthday, Independence Movement Day, and Constitution Day, there were 27 days that corresponded to public holidays during the period covered by the data set. We denoted these dates with a value of "1" if they were public holidays and "0" if they were not, based on the reference date.

Preprocessing and Description of Variables

Among the variables representing individual beds, the reference date, ward and room information, patient occupancy status, bed cleanliness status, and detailed room information were available. Based on the recorded date of bed status, we derived additional variables, such as the reference year, reference month, reference week (week of the year), reference day, and reference day of the week.

Room data were derived from the input information representing the cleanliness status of beds. This variable had 2 possible states, namely, "admittable" and "discharge." If neither of these states was indicated, it implied that a patient was currently hospitalized in the bed. As the status of hospitalized patients was indicated by missing values, we replaced them with the number "1" to indicate the presence of a patient in the bed and "0" otherwise. The sum of all "1" values represented the current number of hospitalized patients. The count of beds in each room indicated the capacity of each room. The target variable BOR was calculated by dividing the number of patients in the room by the room capacity, resulting in a room-specific patient occupancy rate variable. The ward data were subjected to a similar process as that of the room data, with the difference being that we generated ward-specific variables, such as ward capacity and WBOR, using the same approach. The static room data consisted of 14 variables, including the title of the room and the detailed information specific to each room.

For the variables in the ward and room data, we disregarded the units of the features and converted them into numerical values for easy comparison, after which we performed normalization. Regarding the variables representing detailed room information, we converted them to numerical values where "yes" was represented as "1" and "no" was represented as "0."

The final set of variables used in this study was categorized into date, ward, room, and detailed room information. Table 2 provides the detailed descriptions of the variables used in our training, including all the administrative data related to beds that are readily available in the hospital.

The explanation of the classification for generating the data sets for training each model is provided in Table 3. The static features of the detailed room information were combined with the room data set, which has sequence characteristics, to generate a separate data set termed Room+Static.

Table 2. Description of variables by category.

Variable	Type	Description
Date		
Year	3 categories	Reference year for bed status
Month	12 categories	Reference month for bed status
Week	53 categories	Reference week for bed status
Day	31 categories	Reference day for bed status
Weekday	7 categories	Reference day of the week for bed status
Holiday	2 categories	Holiday status
Ward		
Ward abbreviation	57 categories	Abbreviations for entire ward names
Ward capacity	Numeric	Number of available ward beds
Ward bed capacity	Numeric	Number of patients currently admitted to the ward
Ward occupancy rate	Numeric	Ward bed capacity divided by ward capacity
Room		
Room abbreviation	1411 categories	Abbreviations for entire room names
Room capacity	Numeric	Number of available room beds
Room bed capacity	Numeric	Number of patients currently admitted to the room
Room occupancy rate	Numeric	Room bed capacity divided by room capacity
Room static feature		
Room code	34 categories	Room grade code
Nuclear	2 categories (N[a]/Y[b])	Nuclear medicine room availability
Sterile	2 categories (N/Y)	Sterile room availability
Isolation	2 categories (N/Y)	Isolation room availability
EEG[c] testing	2 categories (N/Y)	EEG testing room availability
Observation	2 categories (N/Y)	Observation room availability
Kidney	2 categories (N/Y)	Kidney transplant room availability
Liver	2 categories (N/Y)	Liver transplant room availability
Sub-ICU[d]	2 categories (N/Y)	Sub-ICU room availability
Special	2 categories (N/Y)	Special room availability
Small single	2 categories (N/Y)	Small single room availability
Short-term	2 categories (N/Y)	Short-term room availability
Psy-double	2 categories (N/Y)	Psychiatry department double room availability
Psy-open	2 categories (N/Y)	Psychiatry department open room availability

[a]N: No.

[b]Y: Yes.

[c]EEG: electroencephalogram.

[d]ICU: intensive care unit.

Table 3. Data set classification and included variables.

Data set	Variables
Ward data set	Ward abbreviation, year, month, week, day, weekday, holiday, ward capacity, ward bed capacity, and ward occupancy rate
Room data set	Room abbreviation, year, month, week, day, weekday, holiday, room capacity, room bed capacity, and room occupancy rate
Static data set	14 static variables related to detailed room information
Room+Static data set	Room abbreviation, year, month, week, day, weekday, holiday, room capacity, room bed capacity, 14 static variables related to detailed room information, and room occupancy rate

Separation

Each data set was split into training, validation, and test sets for training and evaluation of the model. The training set consisted of 32,153 rows (67.8%), with data from May 27, 2020, to December 2021. The validation set, used for parameter tuning, included 7085 rows (15.0%), with data from January to June 2022. Finally, the test set comprised 8208 rows (17.2%), with data from July 2022 to November 21, 2022.

DL Algorithms

We used various DL algorithms for in-depth learning. In the following subsections, we will provide explanations for each model algorithm used in our research.

LSTM Network

RNN [26] is a simple algorithm that passes information from previous steps to the current step, allowing it to iterate and process sequential data. However, it encounters difficulties in handling long-term dependencies, such as those found in time-series data, owing to the vanishing gradient problem. To address this issue, LSTM [27] was developed. LSTM excels in handling sequence data and is commonly used in natural language processing, machine translation, and time-series data analysis. LSTM consists of an input gate, output gate, and forget gate. The "cell state," is carefully controlled by each gate to determine whether the memory should be retained or forgotten for the next time step.

Bi-LSTM Network

Although RNN and LSTM possess the ability to remember previous data, they have a limitation in that their results are primarily based on immediate past patterns because the input is processed in a sequential order. This limitation can be overcome through a network architecture known as Bi-LSTM [28]. Bi-LSTM allows end-to-end learning, minimizing the loss on the output and simultaneously training all parameters. It also has the advantage of performing well even with long data sequences. Because of its suitability for models that require knowledge of dependencies from both the past and future, such as LSTM-based time-series prediction, we additionally selected Bi-LSTM as the base model.

Attention Mechanism

Attention mechanism [29,30] refers to the process of incorporating the encoder's outputs into the decoder at each time step of predicting the output sequence. Rather than considering the entire input sequence, it focuses more on the relevant components that are related to the predicted output, allowing the model to focus on important areas. This mechanism helps minimize information loss in data sets with long sequences, enabling better learning and improving the model's performance. It has been widely used in areas such as text translation and speech recognition. Nevertheless, as it is still based on RNN models, it has the drawbacks of slower speed and not being completely free from information loss issues.

Combining Static and Dynamic Features

Data can exhibit different characteristics even at the same time. For instance, in data collected at 1-hour intervals for each hospital bed, we can distinguish between "dynamic data," which include features that change over time, such as the bed condition, date, and patient occupancy, and "static data," which consist of information that remains constant, such as the ward and room number.

DL allows us to use all the available information for prediction. Therefore, for predicting the RBOR, we investigated an approach that combines dynamic and static data using an LSTM-based method [31]. This approach demonstrated better performance than LSTM alone [32]. Our approach involves adding a layer that incorporates static data as an input to the existing room occupancy prediction model.

Model Architecture

Base Model

Our objective was to predict the intermediate-term occupancy rates of wards and rooms within the hospital to contribute to hospital operation planning. Bi-LSTM was chosen as the base model owing to its improved predictive performance compared with the traditional LSTM model. However, to quantitatively compare these models, we conducted a comparison of the results for each model (6 for each, with a total of 12 models).

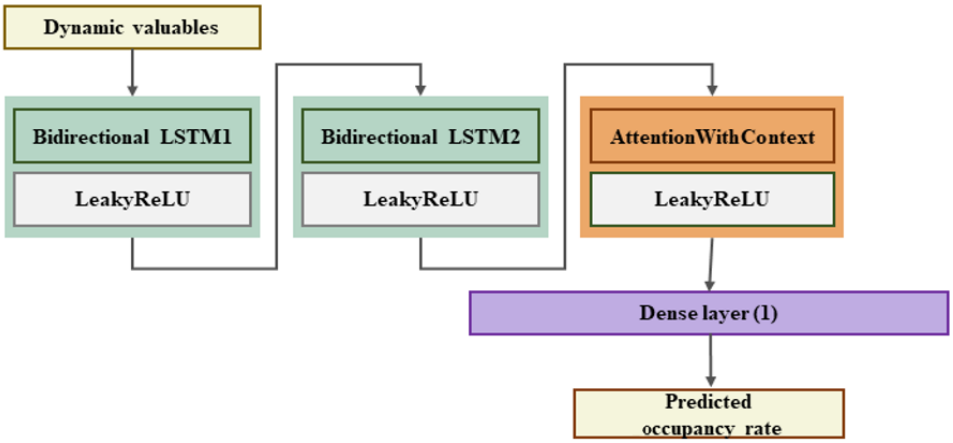
A typical LSTM model processes data sequentially, considering only the information from the past up to the current time step. However, Bi-LSTM, by simultaneously processing data in both forward and backward directions, has a unique feature that allows it to leverage both current and future information for predictions. This bidirectionality helps the model effectively learn temporal dependencies and intricate patterns. However, despite these advantages, Bi-LSTM comes with the trade-off of doubling the number of model parameters, resulting in increased computational costs for training and prediction. While a more complex model can better adapt to the training data, there is an increased risk of overfitting, especially with small

data sets. Nevertheless, the reason for choosing Bi-LSTM for tasks like predicting BORs in hospitals, involving time-series data, lies in its ability to harness the power of bidirectional information. Bi-LSTM processes input data from both past and future directions simultaneously, enabling it to effectively incorporate future information into current predictions. This proves beneficial for handling complex patterns in long time-series data [28].

Moreover, we have enhanced the performance of our models by adding an attention layer to Bi-LSTM. The attention layer assigns higher weights to features that exert a significant impact on the prediction, allowing the model to focus on relevant information and gather necessary input features. This helps improve the accuracy of the prediction. Furthermore, the attention layer reduces the amount of information processed, resulting in improved computational efficiency. Ultimately, this contributes toward enhancing the overall performance of the model.

The window length of the input sequence was divided into 3 different intervals, namely, 3, 7, and 30 days. The WBOR model was trained on sequences with a window length of 7 and 30 days, whereas the RBOR model was trained on sequences with a window length of 3 and 7 days. The first layer of our model consisted of Bi-LSTM, which was followed by the leaky rectified linear unit (LeakyReLU) activation function. LeakyReLU is a linear function that has a small gradient for negative input values, similar to ReLU. It helps the model converge faster. After applying this process once again, the AttentionWithContext layer was applied, which focuses on important components of input sequence data and transforms outputs obtained from the previous layer. After applying the activation function again, a dense layer with 1 neuron was added for generating the final output. The sigmoid function was used to limit the output values between 0 and 1. Finally, our model was compiled using the MSE loss function, Adam optimizer, and MAE metric. The parameters for each layer were selected based on accumulated experience through research. Figure 2 visually represents the above-described structure.

Figure 2. Base bidirectional long short-term memory (Bi-LSTM) model architecture. LeakyReLU: leaky rectified linear unit; LSTM: long short-term memory.



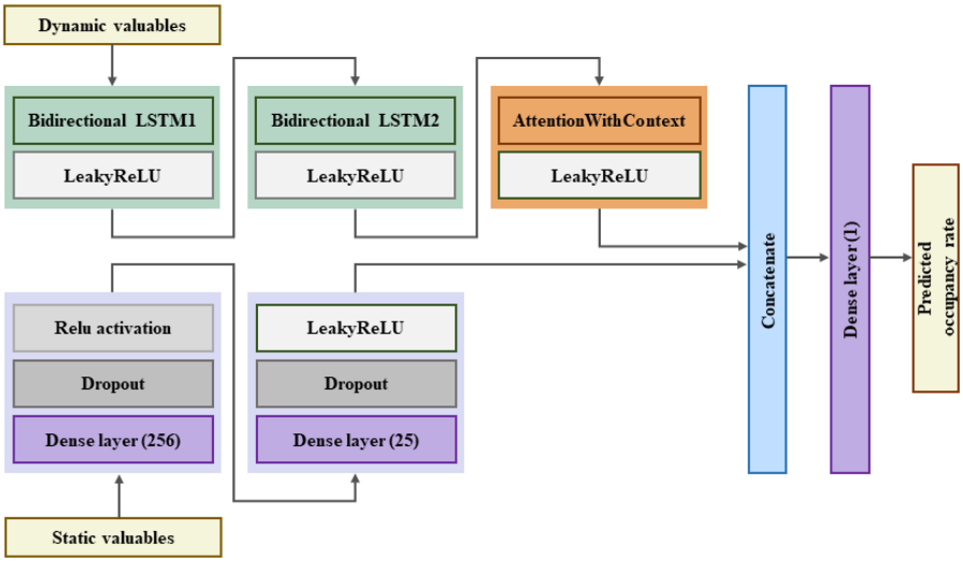
Combining Dynamic and Static Data Using the DL Model

The accumulated bed data, which were collected on a time basis, were divided into dynamic and static data of the rooms, which were then inputted separately. To improve the performance of the BOR prediction model, we designed different DL architectures for the characteristics of these 2 types of data.

We first used a base model based on LSTM and Bi-LSTM to learn the time-series data and then focused the model's attention using the dense layer to process fixed-size inputs. To prevent overfitting, we applied the dropout function to randomly deactivate neurons in 2 dense layers. The hidden states of the 2 networks were combined, and the resulting output was passed to a single layer, combining the time dynamic and static data.

Finally, the hidden states of the 2 networks were combined, and the combined result was passed to a single layer to effectively integrate the dynamic and static data. This allowed us to use the information from both the dynamic and static data for BOR prediction. This architecture is illustrated in Figure 3.

Figure 3. Bidirectional long short-term memory (Bi-LSTM) model architecture combining static and dynamic variables. LeakyReLU: leaky rectified linear unit; LSTM: long short-term memory.



Hyperparameter Tuning

One of the fundamental methods to enhance the performance of artificial intelligence (AI) learning models is the use of hyperparameter tuning. Hyperparameters are parameters passed to the model to modify or adjust the learning process. While hyperparameter tuning may rely on the experience of researchers, there are also functionalities that automatically search for hyperparameters, taking into account the diversity of model structures.

Various methods for search optimization have been proposed [33,34], but we implemented our models using the Keras library. By leveraging Keras Tuner, we automatically searched for the optimal combinations of units and learning rates for each model, contributing to the improvement of their performance.

Time Series Cross-Validation

Time-series data exhibit temporal dependencies between data points, making it crucial to consider these characteristics when validating a model. Commonly used K-fold cross-validation is effective for evaluating models on general data sets [35], providing effectiveness in preventing overfitting and enhancing generalizability by dividing the data into multiple subsets [36,37]. However, for time-series data, shuffling the data randomly is not appropriate owing to the inherent sequential dependency of the observations.

Time series cross-validation is a method that preserves this temporal dependence while dividing the data [38]. It involves splitting the entire hospital bed data set into 5 periods, conducting training and validation for each period, and repeating this process as the periods shift. This approach is particularly effective when observations in the dynamic data set, such as hospital bed data recorded at 1-hour intervals, play a crucial role in predicting future values based on past observations.

Shuffling data randomly using K-fold may disrupt the temporal continuity, leading to inadequate reflection of past and future observations. Therefore, time series cross-validation sequentially partitions the data, ensuring the temporal flow is maintained, and proves to be more effective in evaluating the model's performance. This method enables the model to make more accurate predictions of future occupancy based on past trends.

Evaluation

We selected various metrics to evaluate the performance of time-series data predictions. Among them, MAE represents the absolute difference between the model's predicted values and the actual BOR. We also considered MSE, which is sensitive to outliers. Moreover, to address the limitations of MSE and provide a penalty for large errors, we opted for RMSE. We also used the R^2 score to measure the correlation between the predicted and actual values.

MAE is a commonly used metric to evaluate the performance of time-series prediction models. MAE is intuitive and easy to calculate, making it widely used in practice. Because MAE uses absolute values, it is less sensitive to outliers in the occupancy rate values for specific dates. MAE is calculated using the following formula:

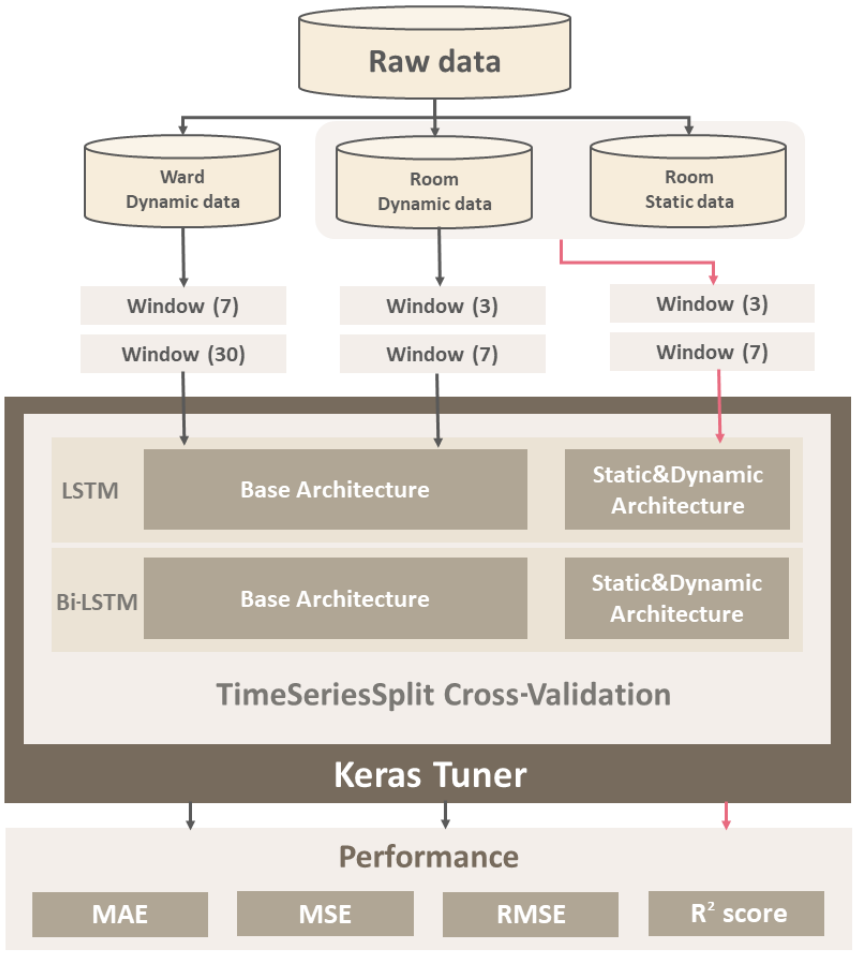
MSE is a metric that evaluates the magnitude of errors by squaring the differences between the predicted and actual values and then taking the average. It is calculated using the following formula:

RMSE is used to address the limitations of MSE where the error scales as a square, providing a more intuitive understanding of the error magnitude between the predicted and actual values. It penalizes large errors, making it less sensitive to outliers. RMSE is calculated using the following formula:

The R^2 score is used to measure the explanatory potential of the prediction model, and it is calculated using the following formula:

Here, SSR represents the sum of squared differences between the predicted and actual values, and SST represents the sum of squared differences between the actual values and the mean value of actual values. Figure 4 shows the prediction method and overall flow in this study.

Figure 4. Overall flow in this study. Bi-LSTM: bidirectional long short-term memory; LSTM: long short-term memory; MAE: mean absolute error; MSE: mean square error; RMSE: root mean square error.



Results

We used 2 DL models, LSTM and Bi-LSTM, and compared the performance of 12 different prediction models. These models have been denoted as ward 7 days (W7D), ward 30 days (W30D), room 3 days (R3D), room 7 days (R7D), room static 3 days (RS3D), and room static 7 days (RS7D). Using Keras Tuner, we adjusted the hyperparameters of the models and subsequently validated the models through a 5-fold time series cross-validation.

The prediction performances of the models for WBOR and RBOR were compared, which showed that they were more accurate at predicting WBOR, with MAE values of 0.06 to 0.07. The W7D model based on Bi-LSTM, which used 7 days of ward data to predict the next day's ward occupancy, had a MAE value of 0.067, MSE value of 0.009, and RMSE value of 0.094, showing high accuracy. The R^2 score was also 0.544, which was approximately 0.240 higher than that of the W30D model (0.304), indicating that the variables in that model explained occupancy reasonably well.

We next compared the performances of the 8 models for RBOR prediction, and among them, the RS7D model based on Bi-LSTM, which was trained on a 7-day time step by integrating static and dynamic data, showed the best performance. It achieved a MAE value of 0.129, MSE value of 0.050, RMSE value of 0.227, and R^2 score of 0.260. In particular, the R^2 score outperformed that of the R3D model by 0.014. These data are summarized in Table 4. Regarding the WBOR prediction model, the model with a shorter training unit, W7D, demonstrated better performance. However, regarding the RBOR prediction model, the model with a longer training unit of 7 days, which incorporated detailed room-specific information, exhibited slightly higher performance than the model with a shorter

training unit of 3 days. The model with the added room-specific information still demonstrated superior performance overall.

We visualized the predicted and actual occupancy for Bi-LSTM models and investigated the occupancy trends since July 2022 on our test data set. First, we selected a specific ward in W7D to demonstrate the change in the WBOR over 2 months. The right panel of Figure 5 shows the WBOR change over 5 months from July 2022 in W30D. The blue line represents the actual occupancy value, and the red line represents the predicted occupancy value by the model. This provides an at-a-glance view of the overall predicted occupancy level for each month and allows hospital staff to observe trends to obtain a rough understanding of the WBOR.

Figure 6 shows graphs of occupancy rate values for a randomized specific room, displaying the predicted and actual values for the 4 RBOR prediction models, with 2 graphs for each model. The left graph shows the occupancy rate change over 5 months from July to November 2022, and the right graph shows the occupancy rate for the months of July and August, providing a detailed view of the RBOR. By examining the trends of the predicted and actual values for the 4 models in this period for a specific room, we can observe that the models maintain a similar trend to the actual occupancy rate.

Table 4. Performances of the occupancy prediction models.

Model and fold	MAE[a]		MSE[b]		RMSE[c]		R^2 score	
	LSTM[d]	Bi-LSTM[e]	LSTM	Bi-LSTM	LSTM	Bi-LSTM	LSTM	Bi-LSTM
Ward								
W30D[f]								
1	0.081	0.097	0.014	0.015	0.117	0.121	0.040	−0.081
2	0.074	0.064	0.011	0.007	0.107	0.085	0.106	0.430
3	0.118	0.109	0.031	0.025	0.175	0.161	−0.130	0.086
4	0.150	0.087	0.033	0.013	0.182	0.113	−0.572	0.399
5	0.087	0.061	0.019	0.008	0.139	0.089	0.212	0.678
Mean	0.102	0.084	0.021	0.014	0.144	0.114	−0.068	0.304
W7D[g]								
1	0.071	0.063	0.011	0.007	0.103	0.086	0.263	0.479
2	0.067	0.054	0.009	0.005	0.094	0.071	0.302	0.606
3	0.119	0.091	0.033	0.016	0.183	0.126	−0.241	0.408
4	0.116	0.068	0.021	0.009	0.145	0.098	−0.009	0.537
5	0.083	0.060	0.015	0.007	0.123	0.087	0.380	0.690
Mean	0.091	0.067	0.018	0.009	0.130	0.094	0.139	0.544
Room								
R7D[h]								
1	0.120	0.111	0.057	0.045	0.238	0.212	0.026	0.226
2	0.127	0.108	0.057	0.047	0.238	0.216	0.054	0.222
3	0.190	0.148	0.167	0.072	0.327	0.269	0.018	0.336
4	0.209	0.162	0.068	0.055	0.261	0.234	−0.089	0.125
5	0.158	0.124	0.069	0.048	0.263	0.220	0.102	0.370
Mean	0.161	0.131	0.071	0.053	0.265	0.230	0.022	0.256
R3D[i]								
1	0.134	0.115	0.058	0.045	0.242	0.212	0.001	0.229
2	0.130	0.097	0.060	0.048	0.245	0.220	0.006	0.195
3	0.178	0.147	0.118	0.080	0.344	0.283	−0.084	0.266
4	0.210	0.204	0.078	0.075	0.280	0.275	−0.247	−0.201
5	0.161	0.120	0.064	0.048	0.254	0.220	0.168	0.377
Mean	0.163	0.167	0.076	0.059	0.273	0.242	−0.031	0.173
RS7D[j]								
1	0.147	0.114	0.057	0.045	0.238	0.212	0.027	0.228
2	0.151	0.099	0.057	0.046	0.240	0.215	0.042	0.227
3	0.216	0.160	0.104	0.063	0.322	0.267	0.048	0.260
4	0.194	0.152	0.064	0.050	0.252	0.224	−0.016	0.198
5	0.181	0.120	0.068	0.047	0.261	0.217	0.112	0.385
Mean	0.178	0.129	0.070	0.050	0.262	0.227	0.043	0.260
RS3D[k]								
1	0.109	0.116	0.056	0.046	0.237	0.215	0.039	0.213

Model and fold	MAE[a]		MSE[b]		RMSE[c]		R^2 score	
	LSTM[d]	Bi-LSTM[e]	LSTM	Bi-LSTM	LSTM	Bi-LSTM	LSTM	Bi-LSTM
2	0.118	0.092	0.061	0.048	0.246	0.219	−0.009	0.203
3	0.182	0.160	0.116	0.090	0.340	0.300	−0.062	0.172
4	0.278	0.191	0.152	0.065	0.389	0.255	−1.410	−0.039
5	0.159	0.116	0.074	0.047	0.272	0.218	0.043	0.387
Mean	0.169	0.135	0.092	0.059	0.297	0.241	−0.028	0.187

[a]MAE: mean absolute error.
[b]MSE: mean square error.
[c]RMSE: root mean square error.
[d]LSTM: long short-term memory.
[e]Bi-LSTM: bidirectional long short-term memory.
[f]W30D: ward 30 days.
[g]W7D: ward 7 days.
[h]R7D: room 7 days.
[i]R3D: room 3 days.
[j]RS7D: room static 7 days.
[k]RS3D: room static 3 days.

Figure 5. Examples of the predicted and actual bed occupancy rates for the 2-month period from July to August 2022 for ward 7 days and the 5-month period from July to November 2022 for ward 30 days.

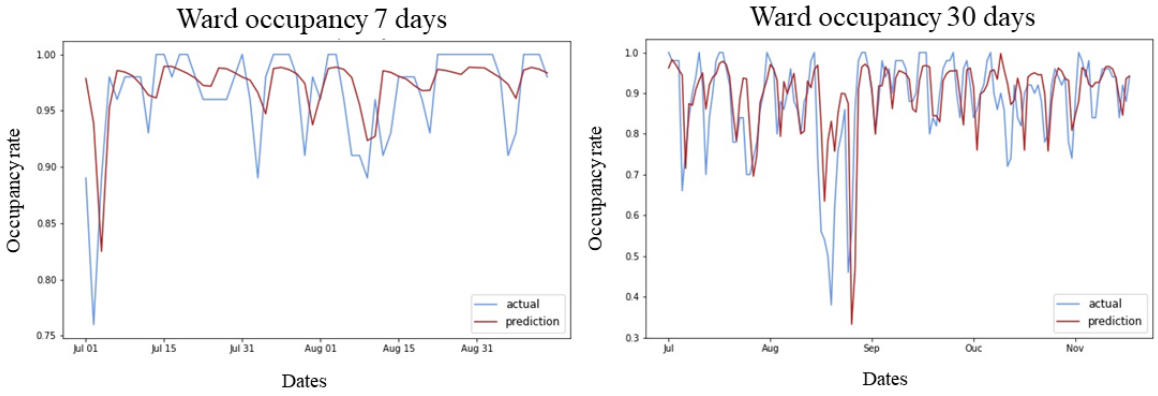
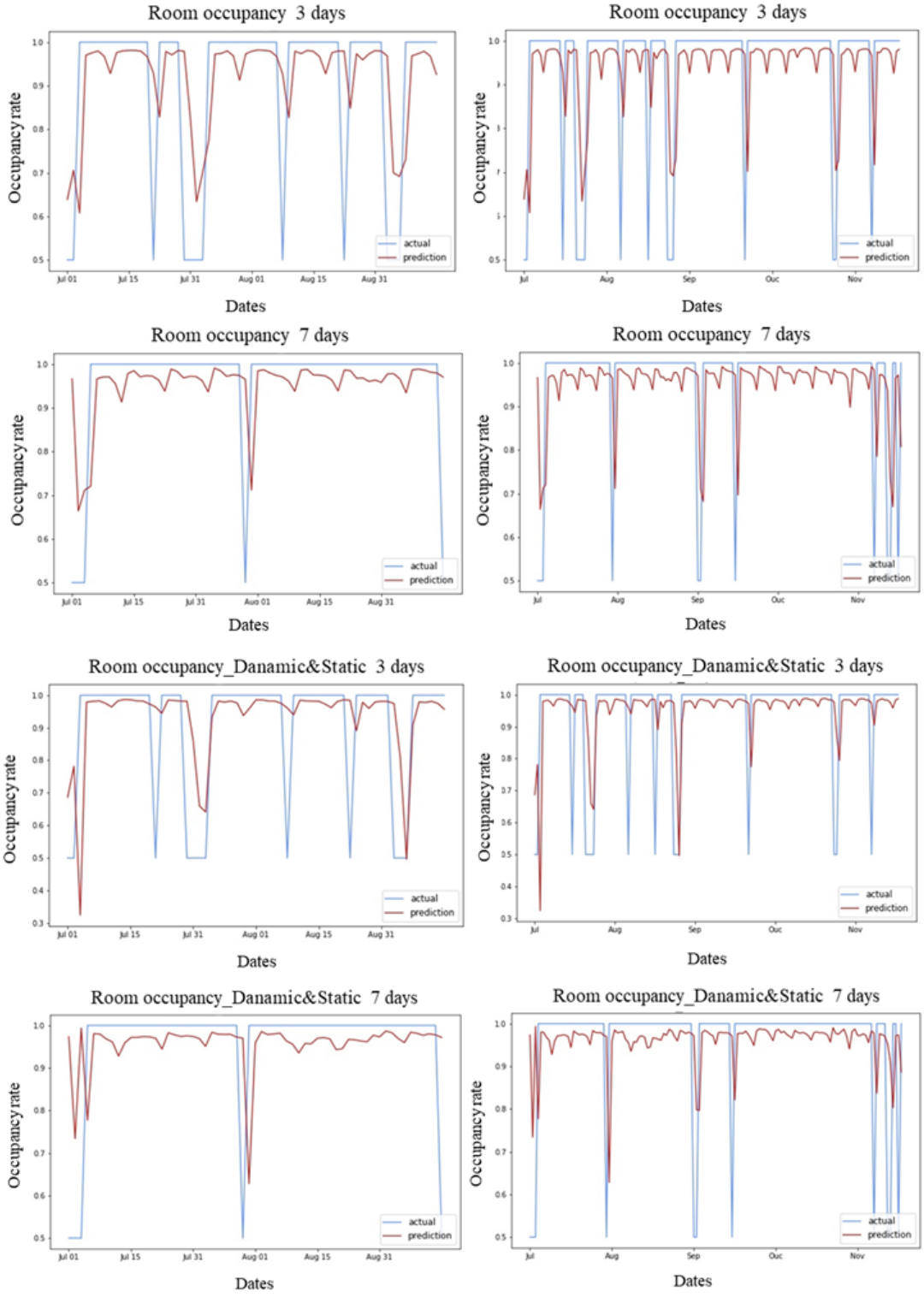


Figure 6. Examples of the predicted and actual bed occupancy rates for the 2-month period from July to August 2022 and the 5-month period from July to November 2022.



Discussion

Principal Findings

The entire data set of this study consisted of administrative data collected at AMC at an hourly interval for each ward from May 27, 2020, to November 21, 2022. To improve the hospital's challenges, we developed a model to predict the occupancy rate of wards and rooms. Our aim was to contribute toward administrative and financial planning for bed management within the hospital.

During the specified period, we compared the results of using DL models to predict the overall BOR for each ward and individual rooms. In the case of WBOR prediction, the MAE of the 7-day window model based on Bi-LSTM was approximately 0.067, demonstrating a remarkably close prediction to the occupancy compared with that of the 30-day

window model based on LSTM, with a difference of approximately 0.035. Furthermore, the MSE and RMSE were 0.009 and 0.094, respectively, indicating high accuracy in the predictions. Moreover, the R^2 score of 0.544 indicated that the model had better explanatory potential than the average. For the individual RBOR prediction, among the 8 models, the RS7D model based on Bi-LSTM performed the best, exhibiting a MAE of approximately 0.129, which was remarkably lower than that of the other models. Moreover, the MSE and RMSE were significantly lower than those of the RBOR models, with differences of 0.042 and 0.07, respectively. The R^2 score of 0.260 indicated that it had higher explanatory potential than the RS3D models based on LSTM, with the value being higher by 0.291.

Finally, we visualized the predicted and actual values on a graph for a specific period and observed that each model captured the trend of the actual BOR quite well. Although the models were less accurate in predicting low occupancy periods, they followed the general trend closely. Overall, these findings demonstrate that our DL models effectively predicted BORs for both wards and individual rooms, with certain models demonstrating superior performance in different scenarios.

Strengths and Limitations

Although the models in this study demonstrated good performance in following the trends of BORs and achieved good results, there were several limitations in this research. First, there were limitations in the data. Although we used administrative data and detailed room information available from the hospital to enable the models to capture occupancy trends, the relationship between the variables and the model's explanatory potential showed room for improvement, as indicated by the R^2 score. To achieve higher prediction accuracy, it would be beneficial to incorporate diverse data sources and real-time updated information.

Second, there was variability in external factors. Hospital BORs are heavily influenced by external environmental factors. Sudden events, such as environmental factors and outbreaks of infectious diseases like COVID-19, can render accurate prediction of bed occupancy challenging [18,32]. Furthermore, seasonal effects and accidents can increase the number of patients. Sufficient collection of long-term data on these external factors would be necessary, but such uncertainties can reduce the accuracy of predictions.

Despite these limitations, our study demonstrated a significant level of adherence to trends in the prediction of individual ward and room occupancy. More detailed variables and a longer period of data accumulation would be required to predict the specific number of beds.

Conclusion

We presented models that can predict the occupancy rates of wards and individual hospital rooms using artificial neural networks based on time-series data. The predicted results of these models demonstrated a high level of accuracy in capturing the future trends of the BOR. In particular, we presented 8 RBOR models with structure and window changes to compare their performance and found that the RS7D model showed the best performance. Our results can be implemented as a web application on hospital online dashboards, as depicted in Figure 1 [25]. In fact, Johns Hopkins University has been applying these methods in their command center to monitor hospital capacity and achieve effectiveness in patient management planning [39].

Furthermore, predicting BORs supports patient admission and discharge planning, helping to alleviate overcrowding in emergency departments and reduce patient waiting times. Staff members can effectively schedule patient admission and discharge, and minimize waiting times by understanding the BOR, providing urgent treatment to emergency patients. Moreover, providing appropriate information to patients waiting in the emergency department can increase patient satisfaction and facilitate efficient transition to hospital admission [40,41]. By applying AI models that combine BOR prediction, which contributes toward reducing emergency department waiting times with individual patient admission and discharge prediction, hospitals can achieve resource optimization and cost savings, resulting in improved patient satisfaction.

Acknowledgments

This work was supported by a Korea Medical Device Development Fund grant funded by the Korean government (the Ministry of Science and ICT; the Ministry of Trade, Industry and Energy; the Ministry of Health & Welfare, Republic of Korea; the Ministry of Food and Drug Safety) (project number: 1711195603, RS-2020-KD000097, 50%) and by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0026).

Conflicts of Interest

None declared.

References

1. Reuben DB, Cassel CK. Physician stewardship of health care in an era of finite resources. JAMA 2011 Jul 27;306(4):430-431. [doi: 10.1001/jama.2011.999] [Medline: 21791692]
2. National Health Expenditure Projections 2011-2021. Centers for Medicare and Medicaid Services. URL: https://www.cms.gov/files/document/forecastsummaryandtablespdf [accessed 2024-02-21]

3. The world health report 2000 - Health systems: improving performance. World Health Organisation. URL: https://cdn.who.int/media/docs/default-source/health-financing/whr-2000.pdf?sfvrsn=95d8b803_1&download=true [accessed 2024-02-21]
4. Kabene SM, Orchard C, Howard JM, Soriano MA, Leduc R. The importance of human resources management in health care: a global context. Hum Resour Health 2006 Jul 27;4(1):20 [FREE Full text] [doi: 10.1186/1478-4491-4-20] [Medline: 16872531]
5. Page K, Barnett AG, Graves N. What is a hospital bed day worth? A contingent valuation study of hospital Chief Executive Officers. BMC Health Serv Res 2017 Feb 14;17(1):137 [FREE Full text] [doi: 10.1186/s12913-017-2079-5] [Medline: 28196489]
6. Keegan AD. Hospital bed occupancy: more than queuing for a bed. Med J Aust 2010 Sep 06;193(5):291-293. [doi: 10.5694/j.1326-5377.2010.tb03910.x] [Medline: 20819049]
7. Kaier K, Mutters N, Frank U. Bed occupancy rates and hospital-acquired infections--should beds be kept empty? Clin Microbiol Infect 2012 Oct;18(10):941-945 [FREE Full text] [doi: 10.1111/j.1469-0691.2012.03956.x] [Medline: 22757765]
8. Anderson D. The impact of resource management on hospital efficiency and quality of care. University of Maryland. 2013. URL: https://api.drum.lib.umd.edu/server/api/core/bitstreams/7ec54849-e2d2-449b-9a9a-f506b429834b/content [accessed 2024-02-21]
9. Kutafina E, Bechtold I, Kabino K, Jonas SM. Recursive neural networks in hospital bed occupancy forecasting. BMC Med Inform Decis Mak 2019 Mar 07;19(1):39 [FREE Full text] [doi: 10.1186/s12911-019-0776-1] [Medline: 30845940]
10. Baas S, Dijkstra S, Braaksma A, van Rooij P, Snijders FJ, Tiemessen L, et al. Real-time forecasting of COVID-19 bed occupancy in wards and Intensive Care Units. Health Care Manag Sci 2021 Jun 25;24(2):402-419 [FREE Full text] [doi: 10.1007/s10729-021-09553-5] [Medline: 33768389]
11. Esteban C, Staeck O, Baier S, Yang Y, Tresp V. Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks. 2016 Presented at: 2016 IEEE International Conference on Healthcare Informatics (ICHI); October 4-7, 2016; Chicago, IL p. 93-101. [doi: 10.1109/ICHI.2016.16]
12. Mackay M, Lee M. Using Compartmental Models to Predict Hospital Bed Occupancy. Semantic Scholar. URL: https://www.semanticscholar.org/paper/Using-Compartmental-Models-to-Predict-Hospital-Bed-Mackay-Lee/f2b32e60df7dd80bd48e8ccd0af920134d1452c5?p2df [accessed 2024-02-21]
13. Littig SJ, Isken MW. Short term hospital occupancy prediction. Health Care Manag Sci 2007 Feb 28;10(1):47-66. [doi: 10.1007/s10729-006-9000-9] [Medline: 17323654]
14. Kumar A, Mo J. Models for Bed Occupancy Management of a Hospital in Singapore. In: Proceedings of the 2010 International Conference on Industrial Engineering and Operations Management. 2010 Presented at: 2010 International Conference on Industrial Engineering and Operations Management; January 9-10, 2010; Dhaka, Bangladesh.
15. Seematter-Bagnoud L, Fustinoni S, Dung D, Santos-Eggimann B, Koehn V, Bize R, et al. Comparison of different methods to forecast hospital bed needs. European Geriatric Medicine 2015 Jun;6(3):262-266. [doi: 10.1016/j.eurger.2014.09.004]
16. Farmer RD, Emami J. Models for forecasting hospital bed requirements in the acute sector. J Epidemiol Community Health 1990 Dec 01;44(4):307-312 [FREE Full text] [doi: 10.1136/jech.44.4.307] [Medline: 2277253]
17. Kim K, Lee C, O'Leary KJ, Rosenauer S, Mehrotra S. Predicting Patient Volumes in Hospital Medicine: A Comparative Study of Different Time Series Forecasting Methods. Northwestern University. URL: https://www.mcs.anl.gov/~kibaekkim/ForecastingHospitalMedicine.pdf [accessed 2024-02-21]
18. Rosenbaum L. Facing Covid-19 in Italy - Ethics, Logistics, and Therapeutics on the Epidemic's Front Line. N Engl J Med 2020 May 14;382(20):1873-1875. [doi: 10.1056/NEJMp2005492] [Medline: 32187459]
19. Bouhamed H, Hamdi M, Gargouri R. Covid-19 Patients' Hospital Occupancy Prediction During the Recent Omicron Wave via some Recurrent Deep Learning Architectures. Int. J. Comput. Commun. Control 2022 Mar 14;17(3):4697. [doi: 10.15837/ijccc.2022.3.4697]
20. Bekker R, Uit Het Broek M, Koole G. Modeling COVID-19 hospital admissions and occupancy in the Netherlands. Eur J Oper Res 2023 Jan 01;304(1):207-218 [FREE Full text] [doi: 10.1016/j.ejor.2021.12.044] [Medline: 35013638]
21. Farcomeni A, Maruotti A, Divino F, Jona-Lasinio G, Lovison G. An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. Biom J 2021 Mar 30;63(3):503-513 [FREE Full text] [doi: 10.1002/bimj.202000189] [Medline: 33251604]
22. Caro JJ, Möller J, Santhirapala V, Gill H, Johnston J, El-Boghdadly K, et al. Predicting Hospital Resource Use During COVID-19 Surges: A Simple but Flexible Discretely Integrated Condition Event Simulation of Individual Patient-Hospital Trajectories. Value Health 2021 Nov;24(11):1570-1577 [FREE Full text] [doi: 10.1016/j.jval.2021.05.023] [Medline: 34711356]
23. Schmidt R, Geisler S, Spreckelsen C. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. BMC Med Inform Decis Mak 2013 Jan 07;13:3 [FREE Full text] [doi: 10.1186/1472-6947-13-3] [Medline: 23289448]
24. Hancock WM, Walter PF. The use of computer simulation to develop hospital systems. SIGSIM Simul. Dig 1979 Jul;10(4):28-32. [doi: 10.1145/1102815.1102819]

25. Shahpori R, Gibney N, Guebert N, Hatcher C, Zygun D. An on-line dashboard to facilitate monitoring of provincial ICU bed occupancy in Alberta, Canada. JHA 2013 Oct 10;3(1):47. [doi: 10.5430/jha.v3n1p47]
26. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct 9;323(6088):533-536. [doi: 10.1038/323533a0]
27. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
28. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans. Signal Process 1997;45(11):2673-2681. [doi: 10.1109/78.650093]
29. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv. 2014. URL: https://arxiv.org/abs/1409.0473 [accessed 2024-02-21]
30. Luong MT, Pham H, Manning CD. Effective Approaches to Attention-based Neural Machine Translation. arXiv. 2015. URL: https://arxiv.org/abs/1508.04025 [accessed 2024-02-21]
31. Leontjeva A, Kuzovkin I. Combining Static and Dynamic Features for Multivariate Sequence Classification. 2016 Presented at: 2016 IEEE 3rd International Conference on Data Science and Advanced Analytics (DSAA); October 17-19, 2016; Montreal, QC p. 21-30. [doi: 10.1109/DSAA.2016.10]
32. Vincent J, Creteur J. Ethical aspects of the COVID-19 crisis: How to deal with an overwhelming shortage of acute beds. Eur Heart J Acute Cardiovasc Care 2020 Apr 29;9(3):248-252 [FREE Full text] [doi: 10.1177/2048872620922788] [Medline: 32347745]
33. Vakharia V, Shah M, Nair P, Borade H, Sahlot P, Wankhede V. Estimation of Lithium-ion Battery Discharge Capacity by Integrating Optimized Explainable-AI and Stacked LSTM Model. Batteries 2023 Feb 09;9(2):125. [doi: 10.3390/batteries9020125]
34. Joshi S, Owens JA, Shah S, Munasinghe T. Analysis of Preprocessing Techniques, Keras Tuner, and Transfer Learning on Cloud Street image data. 2021 Presented at: IEEE International Conference on Big Data (Big Data); December 15-18, 2021; Orlando, FL. [doi: 10.1109/BigData52589.2021.9671878]
35. Jung Y. Multiple predicting K-fold cross-validation for model selection. Journal of Nonparametric Statistics 2017 Nov 21;30(1):197-215. [doi: 10.1080/10485252.2017.1404598]
36. Nair P, Vakharia V, Borade H, Shah M, Wankhede V. Predicting Li-Ion Battery Remaining Useful Life: An XDFM-Driven Approach with Explainable AI. Energies 2023 Jul 31;16(15):5725. [doi: 10.3390/en16155725]
37. Seo H, Ahn I, Gwon H, Kang HJ, Kim Y, Cho HN, et al. Prediction of hospitalization and waiting time within 24 hours of emergency department patients with unstructured text data. Health Care Manag Sci 2023 Nov 03:09660-5. [doi: 10.1007/s10729-023-09660-5] [Medline: 37921927]
38. Deng A. Time series cross validation: A theoretical result and finite sample performance. Economics Letters 2023 Dec;233:111369. [doi: 10.1016/j.econlet.2023.111369]
39. Martinez DA, Kane EM, Jalalpour M, Scheulen J, Rupani H, Toteja R, et al. An Electronic Dashboard to Monitor Patient Flow at the Johns Hopkins Hospital: Communication of Key Performance Indicators Using the Donabedian Model. J Med Syst 2018 Jun 18;42(8):133. [doi: 10.1007/s10916-018-0988-4] [Medline: 29915933]
40. Gartner D, Padman R. Machine learning for healthcare behavioural OR: Addressing waiting time perceptions in emergency care. Journal of the Operational Research Society 2019 Apr 15;71(7):1087-1101. [doi: 10.1080/01605682.2019.1571005]
41. Welch SJ. Twenty years of patient satisfaction research applied to the emergency department: a qualitative review. Am J Med Qual 2010 Dec 04;25(1):64-72. [doi: 10.1177/1062860609352536] [Medline: 19966114]

Abbreviations

AI: artificial intelligence
AMC: Asan Medical Center
Bi-LSTM: bidirectional long short-term memory
BOR: bed occupancy rate
DL: deep learning
DNN: deep neural network
LeakyReLU: leaky rectified linear unit
LSTM: long short-term memory
MAE: mean square error
ML: machine learning
R3D: room 3 days
R7D: room 7 days
RBOR: room bed occupancy rate
RMSE: root mean square error
RNN: recurrent neural network
RS3D: room static 3 days

RS7D: room static 7 days
W7D: ward 7 days
W30D: ward 30 days
WBOR: ward bed occupancy rate

Edited by C Lovis; submitted 05.10.23; peer-reviewed by V Vakharia, T Leili; comments to author 10.11.23; revised version received 20.12.23; accepted 16.02.24; published 21.03.24.

Please cite as:
Seo H, Ahn I, Gwon H, Kang H, Kim Y, Choi H, Kim M, Han J, Kee G, Park S, Ko S, Jung H, Kim B, Oh J, Jun TJ, Kim YH
Forecasting Hospital Room and Ward Occupancy Using Static and Dynamic Information Concurrently: Retrospective Single-Center Cohort Study
JMIR Med Inform 2024;12:e53400
URL: https://medinform.jmir.org/2024/1/e53400
doi:10.2196/53400
PMID:38513229

Original Paper

Explainable AI Method for Tinnitus Diagnosis via Neighbor-Augmented Knowledge Graph and Traditional Chinese Medicine: Development and Validation Study

Ziming Yin[1*], Prof Dr; Zhongling Kuang[1*], MSc; Haopeng Zhang[2], MD; Yu Guo[2], MD; Ting Li[1], BEng; Zhengkun Wu[1], BEng; Lihua Wang[2], MD

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China

[2]Department of Otolaryngology, Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Lihua Wang, MD
Department of Otolaryngology
Shanghai Municipal Hospital of Traditional Chinese Medicine
Shanghai University of Traditional Chinese Medicine
274 Zhijiang Middle Road, Jing'an District
Shanghai, 200071
China
Phone: 86 18116013561
Email: lihuahanhan@126.com

Abstract

Background: Tinnitus diagnosis poses a challenge in otolaryngology owing to an extremely complex pathogenesis, lack of effective objectification methods, and factor-affected diagnosis. There is currently a lack of explainable auxiliary diagnostic tools for tinnitus in clinical practice.

Objective: This study aims to develop a diagnostic model using an explainable artificial intelligence (AI) method to address the issue of low accuracy in tinnitus diagnosis.

Methods: In this study, a knowledge graph–based tinnitus diagnostic method was developed by combining clinical medical knowledge with electronic medical records. Electronic medical record data from 1267 patients were integrated with traditional Chinese clinical medical knowledge to construct a tinnitus knowledge graph. Subsequently, weights were introduced, which measured patient similarity in the knowledge graph based on mutual information values. Finally, a collaborative neighbor algorithm was proposed, which scored patient similarity to obtain the recommended diagnosis. We conducted 2 group experiments and 1 case derivation to explore the effectiveness of our models and compared the models with state-of-the-art graph algorithms and other explainable machine learning models.

Results: The experimental results indicate that the method achieved 99.4% accuracy, 98.5% sensitivity, 99.6% specificity, 98.7% precision, 98.6% F_1-score, and 99% area under the receiver operating characteristic curve for the inference of 5 tinnitus subtypes among 253 test patients. Additionally, it demonstrated good interpretability. The topological structure of knowledge graphs provides transparency that can explain the reasons for the similarity between patients.

Conclusions: This method provides doctors with a reliable and explainable diagnostic tool that is expected to improve tinnitus diagnosis accuracy.

(JMIR Med Inform 2024;12:e57678) doi:10.2196/57678

KEYWORDS

knowledge graph; syndrome differentiation; tinnitus; traditional Chinese medicine; explainable; ear; audiology; TCM; algorithm; diagnosis; AI; artificial intelligence

Introduction

Tinnitus is a common refractory disease in the field of otolaryngology, and its diagnosis has always been a cutting-edge research topic in audiology. With changes in the social environment and an accelerated pace of life, an increasing number of patients, particularly among the younger generation, have sought medical assistance for tinnitus as their primary complaint in the last decade. Globally, approximately 14% (95% CI 0.8%-1.6%) of adults are affected by tinnitus [1,2], which can cause stress, anxiety, and depression [3]. Distress and hearing impairment brought on by the disease can affect cognitive abilities and lead to suicidal tendencies in severe cases, greatly affecting the work and daily lives of patients [4].

The pathogenesis of tinnitus is extremely complex and not fully understood. Currently, no effective objectification methods are available. Traditional Chinese medicine (TCM) classifies tinnitus into 5 different syndrome patterns: wind fire attacking internally (WFAI), liver fire bearing upward (LFBU), phlegm fire stagnation internally (PFSI), Qi deficiency of the spleen and stomach (QDSS), and kidney essence deficiency (KED). The diagnosis of tinnitus remains a challenge in medical science because it is influenced by several complex factors [5,6], including individual differences among patients and atypical symptom presentations. Clinical diagnosis relies heavily on the personal knowledge and clinical experience of doctors, thereby introducing subjectivity, uncertainty, and ambiguity. Consequently, achieving a high tinnitus diagnostic accuracy becomes difficult. Therefore, tinnitus diagnosis remains an urgent issue requiring further exploration and resolution by medical researchers.

Previous studies have focused on the use of artificial intelligence (AI) to assist doctors in diagnosing tinnitus and improving diagnostic accuracy. Liu et al [7] proposed a meta-learning method based on lateral perception for cross–data set tinnitus diagnosis. Sun et al [8] used a support vector machine classifier to distinguish between patients with tinnitus and healthy individuals. Shoushtarian et al [9] used a naive Bayes algorithm to classify patients with tinnitus and control groups. Sanders et al [10] used a spiking neural network model to classify patients with tinnitus into 2 groups based on different classification criteria. Manta et al [11] used clinical data and patient features to build a machine learning (ML) model for classifying the degree of tinnitus-related distress in individuals and their ears. Allgaier et al [12] used a gradient-boosting engine to classify transient tinnitus. Rodrigo et al [13] used a decision tree model to identify variables related to the success of internet-based cognitive behavioral therapy for tinnitus. Liu et al [14] used a support vector machine model to explore cortical or subcortical morphological neuroimaging biomarkers that effectively distinguished patients with tinnitus from healthy individuals. Niemann et al [15] proposed a LASSO model to predict the severity of depression in patients with tinnitus. Although previous studies have achieved success using their respective data sets, the developed ML- or deep learning–based methods are entirely data-driven modeling approaches that do not make full use of existing medical knowledge. Models built using such methods are equivalent to “black boxes” for doctors, lack interpretability, and are not conducive to clinical promotion and application.

In this study, the aim is to incorporate clinical medical knowledge into a diagnostic model, enabling the integration of knowledge and data for interpretable results. Knowledge graph–based modeling methods offer solutions to such issues by using a novel knowledge representation format that connects entities and concepts in an objective world using semantic relationships. Such methods offer reasoning and interpretability that are highly sought after by both medical practitioners and academia. Li et al [16] used a knowledge graph to predict diabetic macular edema, overcoming the limitations of traditional ML and data-mining techniques that deal with missing feature values. Zhou et al [17] used 124 medical records to construct a knowledge graph for recommending hypertension medication. Lyu et al [18] created a knowledge graph for diabetic nephropathy diagnosis using patient data. Lin et al [19] extracted knowledge from medical texts and historical prescription data to construct a medical knowledge graph and accurately detect clinical prescription risks. Recently, knowledge graph applications have expanded to TCM; for instance, Yang et al [20] built a knowledge graph to extract medical information from TCM case records. Xie et al [21] constructed a knowledge graph using ancient Chinese medical books to infer symptoms and syndromes. Yang et al [22] used electronic medical records (EMRs) to build a knowledge graph, transforming TCM diagnostic issues into multilabel classification problems. Lan et al [23] integrated knowledge graphs with graph neural networks to introduce graph-based supervised contrastive learning, effectively enabling the classification of TCM texts. However, no previous studies have used knowledge graphs in the complex medical field of tinnitus diagnosis. Therefore, this study focuses on knowledge graph technology to assist doctors in tinnitus diagnosis and improve diagnostic accuracy.

This paper aims to establish a comprehensive knowledge graph in TCM specifically tailored for tinnitus. Leveraging this knowledge graph, we propose a novel method for calculating patient similarity. This method takes into account the weighting of symptom-syndrome type relationships, thereby facilitating the inference of syndrome types in patients with tinnitus according to TCM principles. By implementing this approach, clinicians can increase the accuracy of tinnitus diagnosis within the realm of TCM.

In general, we make several noteworthy contributions as follows:

- We propose a method for tinnitus knowledge graph construction based on heterogeneous patient EMRs and TCM clinical knowledge.
- We introduce weights to measure patient similarity into the tinnitus knowledge graph using a method based on prior probabilities and mutual information values.
- A collaborative neighbor algorithm that uses patient similarity scores to obtain recommended diagnostic results is proposed to assist doctors in understanding the model-generated conclusions, thereby improving the accuracy of tinnitus diagnosis.

Methods

Patients

For this study, we collected the EMRs of 1267 patients with tinnitus who visited the ear, nose, and throat departments of 11 medical institutions in Shanghai, China, from November 2019 to July 2023. The inclusion criteria included (1) tinnitus as the primary complaint and (2) the ability to communicate normally. The exclusion criteria included (1) objective tinnitus, (2) nonotogenic tinnitus caused by factors such as endocrine and blood disorders, (3) tinnitus caused by head or ear trauma, and (4) difficulties in communication or severe psychiatric history that could hinder follow-up compliance. After screening the data for quality, 1265 cases were included for further analysis.

The clinical EMR data set recorded medical data of real patients including the relationship between patient symptoms and disease, which was crucial for disease diagnosis. The data set contained patient information such as age, sex, inducement, medical history, tinnitus sound, accompanying symptoms, tongue coating, pulse condition, TCM syndrome differentiation, and sleep status. Each patient had a clear diagnosis that could be classified into 1 of 5 categories: WFAI, LFBU, PFSI, QDSS, and KED. Statistical data are presented in Figures 1-4 .

Figure 1. Age distribution of different syndromes by sex. KED: kidney essence deficiency; LFBU: liver fire bearing upward; PFSI: phlegm fire stagnation internally; QDSS: Qi deficiency of the spleen and stomach; WFAI: wind fire attacking internally.

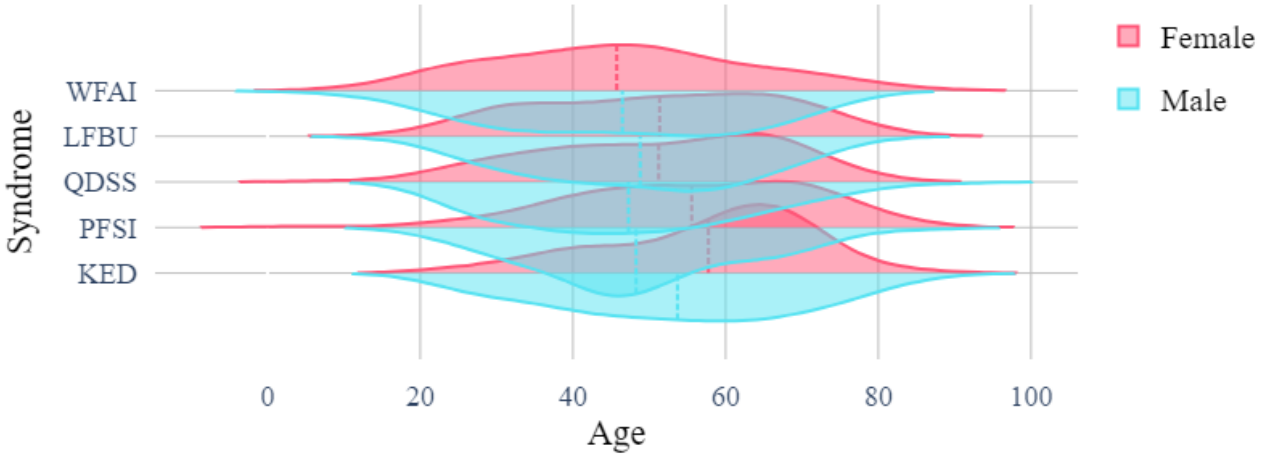


Figure 2. The tongue body distribution of different syndrome types. KED: kidney essence deficiency; LFBU: liver fire bearing upward; PFSI: phlegm fire stagnation internally; QDSS: Qi deficiency of the spleen and stomach; WFAI: wind fire attacking internally.

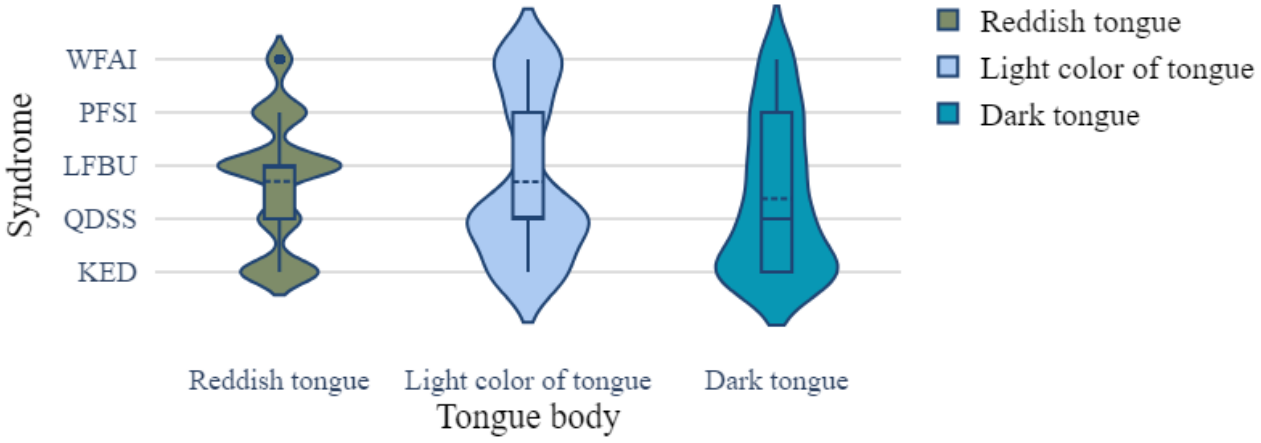


Figure 3. The tongue fur distribution of different syndrome types. KED: kidney essence deficiency; LFBU: liver fire bearing upward; PFSI: phlegm fire stagnation internally; QDSS: Qi deficiency of the spleen and stomach; WFAI: wind fire attacking internally.

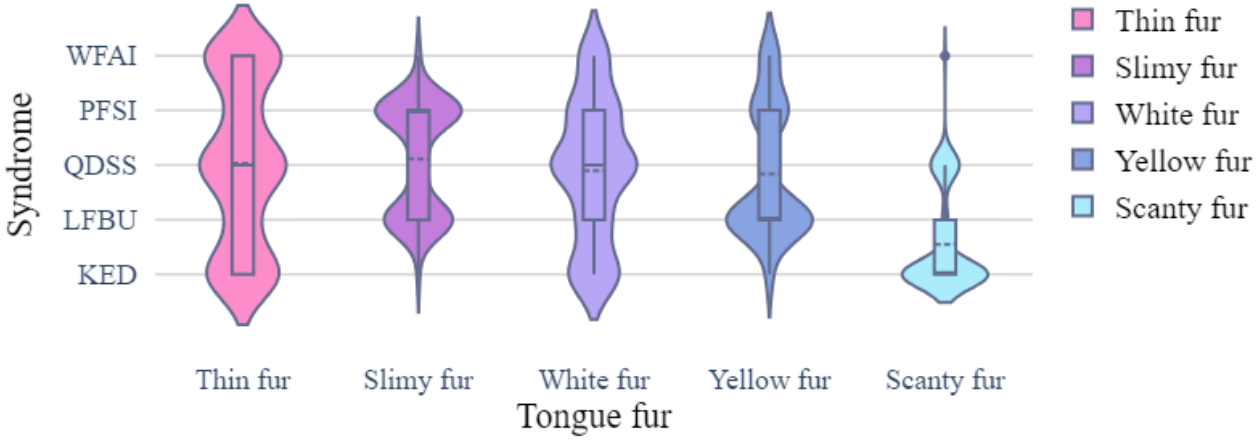
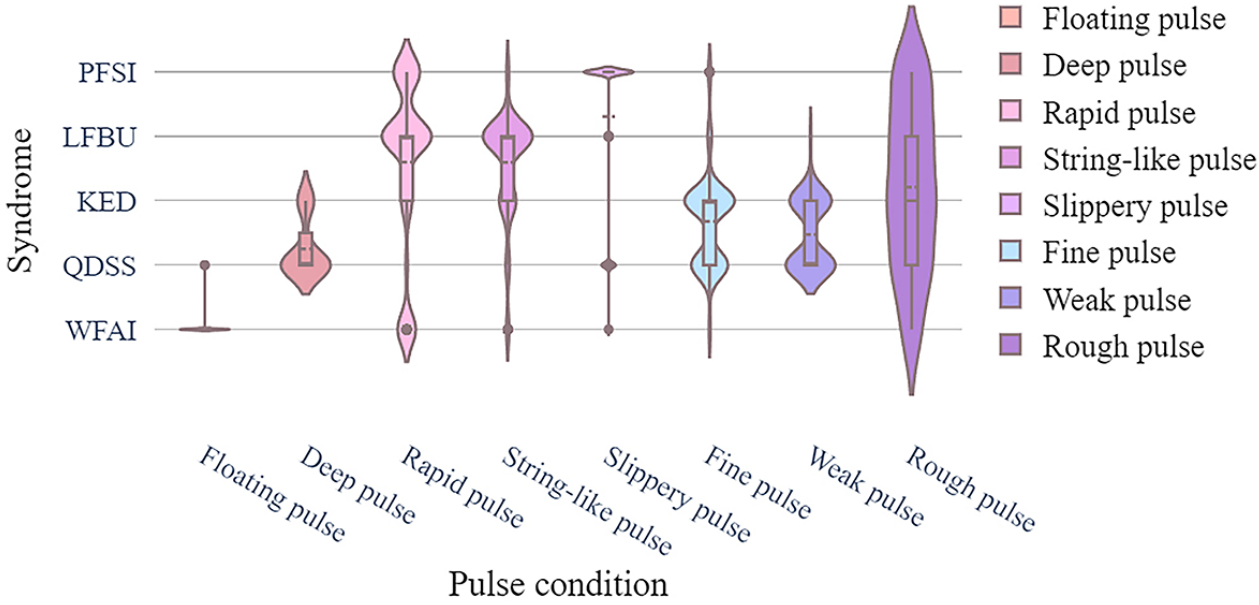


Figure 4. The pulse condition distribution of different syndrome types. KED: kidney essence deficiency; LFBU: liver fire bearing upward; PFSI: phlegm fire stagnation internally; QDSS: Qi deficiency of the spleen and stomach; WFAI: wind fire attacking internally.



Ethical Considerations

This study's protocol was approved by the ethics committee of the Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai, China (2021SHL-KY-70).

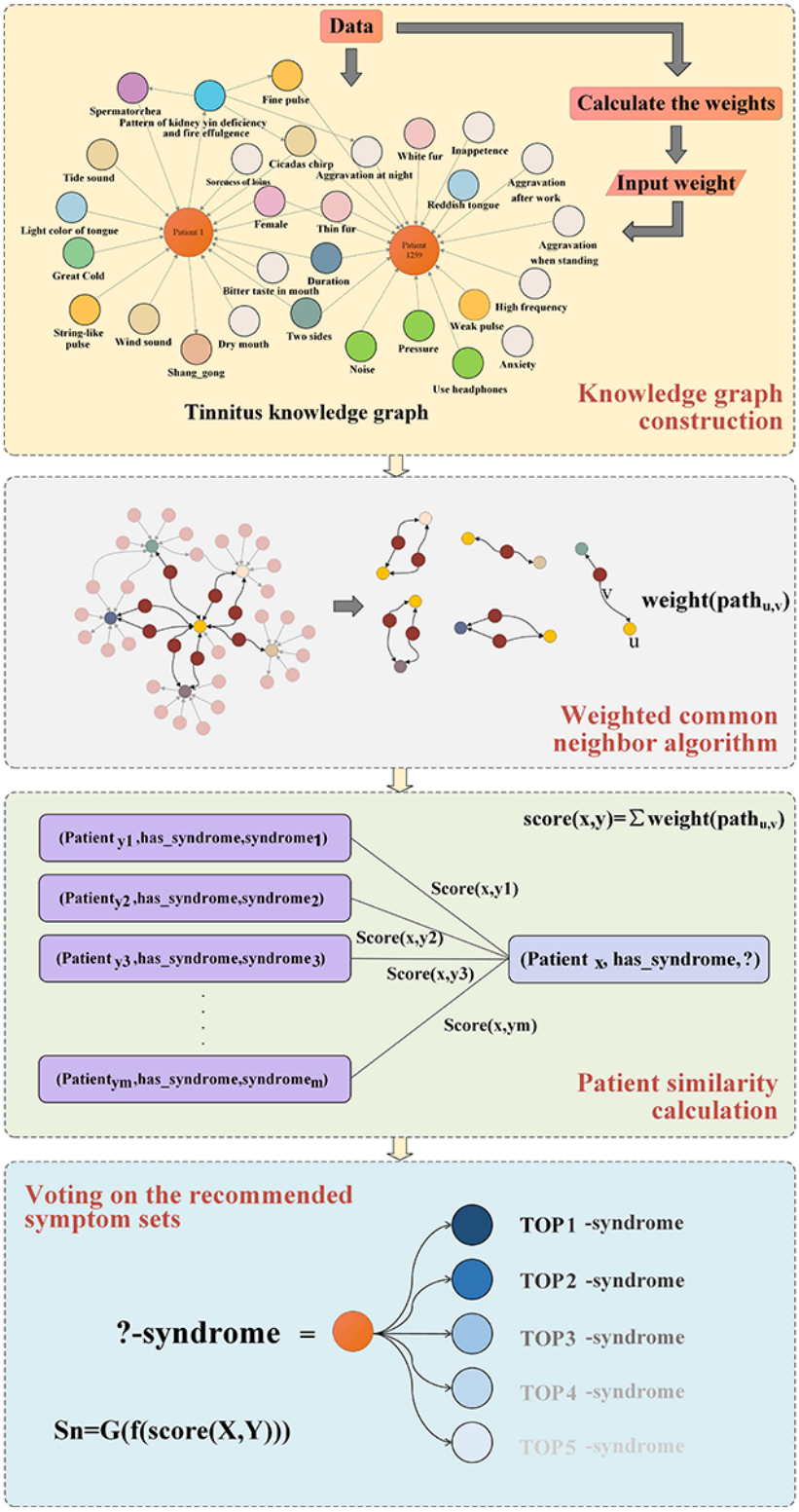
The data was anonymized in order to protect patient privacy. Patients could receive free examinations and treatments throughout the entire process, so no compensation was provided.

Clinical Decision Support for Tinnitus

Overview

To integrate patient EMRs with diagnostic knowledge from TCM textbooks, we constructed a knowledge graph using a combined "top-down" and "bottom-up" approach [24]. First, a patient-centered knowledge graph was developed using EMRs. Then, the knowledge graph was enriched with tinnitus diagnostic knowledge from TCM textbooks. Finally, we used a mutual information–based weight calculation method to enhance the knowledge graph by fusing patient case data with diagnostic knowledge. The resulting knowledge graph simulated the diagnostic reasoning processes of experienced physicians. The entire method consisted of three steps: (1) building a weighted tinnitus knowledge graph, (2) finding and scoring common neighbors, and (3) predicting syndrome patterns based on patient similarity. The overall framework is illustrated in Figure 5.

Figure 5. Overall framework of the proposed method.



Knowledge Graph of Tinnitus Based on Heterogeneous Sources

In response to the diagnostic needs of tinnitus in TCM, the ontology structure of a tinnitus medical knowledge graph should revolve around symptoms, syndrome patterns, diseases, drugs, and treatment methods. For this study, we extracted such common concepts from expert-reviewed EMRs and classic medical textbooks, constructed a conceptual knowledge system, and built a top-level ontology structure. Natural language processing techniques [25] were used to extract entities and relationships from the patient EMRs based on a defined conceptual knowledge system for tinnitus. By applying certain rules and conducting string matching within the text, we extracted 15 and 10 categories of entities and relationships from the 1265 EMR records, respectively. Once the entity types and hierarchy were determined, we embedded the data into the conceptual knowledge system and established a patient-centric

tinnitus knowledge graph in the form of a triple, which maximized the retention of both explicit and implicit diagnostic information.

Furthermore, we enhanced the constructed tinnitus knowledge graph using knowledge extracted from authoritative medical textbooks to supplement tinnitus knowledge information that was not fully expressed in EMRs. Together with the EMR knowledge graph, a complete tinnitus knowledge graph was developed. The knowledge we selected came from 2 classic Chinese medicine textbooks [26,27], from which we extracted basic concepts related to tinnitus including TCM syndromes, prescriptions, Chinese medicinal herbs, and treatment methods to construct the TCM knowledge graph.

Heterogeneous Knowledge Fusion

Redundancy in the entities and relationships extracted from heterogeneous sources was observed owing to the different sources of data and knowledge. Therefore, knowledge fusion was required. First, data normalization and entity alignment were performed to standardize the named entities extracted from multiple data sources. The entities were associated using string-matching and similarity-calculation methods. As entity and attribute texts were relatively short, a lower similarity threshold was more appropriate; therefore, the similarity judgment threshold was set as 0.6 to prevent errors and omissions. The entity similarity calculation results are listed in Table 1. As the knowledge graph was established in Chinese, we calculated the similarity of the Chinese strings.

Then, a matching path was built from the tinnitus ontology–based knowledge graph entity to the EMR-based knowledge graph entity. Patient data were linked to diagnostic knowledge through an ontology. The 2 knowledge graphs were linked by unifying entities with duplicate meanings in the 2 graphs. Manual verification was performed to ensure the accuracy of the knowledge graph. The specific method is illustrated in Figure 6. Finally, the tinnitus knowledge graph consisted of 1247 entities and 9234 relationships.

Table 1. Entity similarity calculation results.

Standardized and ambiguous entities (Chinese)	Similarity
WFAI[a]	
风热外侵证 (wind-heat invasion syndrome)	0.8
风热外犯证 (wind-heat exterior syndrome)	0.6
风热外侵证 (wind-heat exterior assault syndrome)	0.8
LFBU[b]	
肝火上炎证 (liver fire flaming upward syndrome)	0.8
肝热上扰证 (liver heat disturbing upward syndrome)	0.8
肝火上扰清窍证 (liver fire disturbing upward and disturbing clearing orifices syndrome)	0.83
QDSS[c]	
脾胃虚证 (spleen and stomach deficiency syndrome)	0.89
脾胃虚弱证 (spleen and stomach weakness syndrome)	0.8
PFSI[d]	
痰火壅结证 (phlegm-fire concretions syndrome)	0.8
KED[e]	
肾精不足证 (kidney essence insufficiency syndrome)	0.6
肾精亏虚证 (kidney essence deficiency syndrome)	0.8
肾虚精亏证 (kidney deficiency and essence deficiency syndrome)	0.99
肾精亏耗证 (kidney essence consumption syndrome)	0.8

[a]WFAI: wind fire attacking internally.

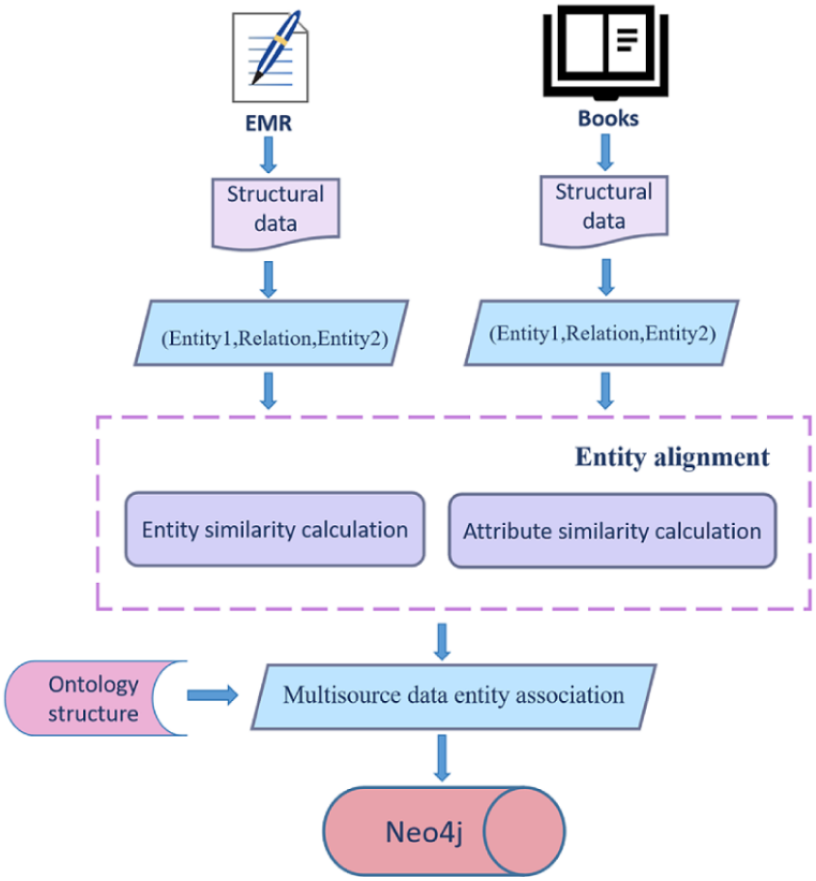
[b]LFBU: liver fire bearing upward.

[c]QDSS: Qi deficiency of the spleen and stomach.

[d]PFSI: phlegm fire stagnation internally.

[e]KED: kidney essence deficiency.

Figure 6. Tinnitus knowledge graph fusion flowchart. EMR: electronic medical record.



Calculation of Knowledge Graph Relationship Weights Based on Mutual Information

Considering the varying importance of different entities for different syndrome patterns, the imbalance in data categories, and the varying amount of information carried by symptoms, the calculation of weights required consideration of entities' importance for diagnostic pattern identification and information content carried by the entities themselves. The data used for weight calculation were derived from real clinical case data used for constructing the knowledge graph. First, the mutual information value (w_{if}) possessed by each entity was obtained using the mutual information method. The obtained value represented the extent to which a variable could acquire diagnostic pattern information.

For a given set of entities $X = \{x_1, x_2, ..., x_n\}$ with corresponding probabilities $P = \{p_1, p_2, ..., p_n\}$, the target variable to be measured was the diagnostic pattern Y. By calculating the overall entropy $H()$, conditional entropy $H(Y|X)$, and mutual information value $Gain(S,x)$, the degree to which the diagnostic pattern was determined based on the entity values or the weight value w_{if} of the entity was calculated. The calculations were performed using equations 1-3.

(1)

(2)

$$w_{if} = Gain(Y,X) = H(Y) - H(Y|X) \quad (3)$$

Further, the feature weights were calculated based on the syndrome patterns under the prior conditions. The probability of each symptom appearing under different syndrome patterns was obtained using statistical methods such as:

$$w_{sd} = p(sym_i|sd_j) \quad (4)$$

where $sym = \{sym_1, sym_2, ..., sym_n\}$ represents the symptom set and $sd = \{sd_1, sd_2, ..., sd_m\}$ represents the diagnostic pattern set. Finally, the edge weight from node u to node v was defined using equation 5.

$$Weight(u,v) = w_{if} + w_{sd} \quad (5)$$

The weights of various symptoms under different syndrome patterns are presented in Table 2.

Table 2. Partial weight value of symptom-syndrome type.

Symptom	Weight
KED[a]	
Spermatorrhea	1.435
Soreness of loins	1.4213
Dreaminess	1.4104
Wake up early in the morning	1.3868
Deficiency and insomnia	1.3856
Aggravation at night	1.167
Cicadas chirp	1.1559
Fine pulse	1.1448
Scanty fur	0.7142
Duration	0.6991
LFBU[b]	
Irritable	1.2376
Restlessness and insomnia	1.1196
Wind sound	1.0271
String-like pulse	1.0056
Tide sound	1.0030
Yellow fur	0.9118
Reddish tongue	0.8992
Duration	0.7036
Dry mouth	0.6855
Bitter taste in mouth	0.6558
PFSI[c]	
Tastelessness	1.1953
Dizziness and heaviness	1.1488
Aural fullness	1.1216
Ear distension	1.0899
Slippery pulse	0.9121
Slimy fur	0.8342
Duration	0.7113
Yellow fur	0.6895
Hearing loss	0.6495
Reddish tongue	0.6440
WFAI[d]	
Cold or rhinitis	1.2089
Tinnitus onset within a month	1.1398
Low voice	1.1398
Thin fur	1.0286
Floating pulse	0.9563
Duration	0.6903
Light color of tongue	0.6664

Symptom	Weight
Yellow fur	0.5082
Hearing loss	0.5032
Dreaminess	0.4993
QDSS[e]	
Feeling emptiness in ear	1.2615
Aggravation after work	1.1813
Aggravation when standing up	1.1562
Fine pulse	1.0782
Duration	0.7370
Thin fur	0.7022
Light color of tongue	0.6745
Anxiety	0.6596
Hearing loss	0.6444
Dreaminess	0.4865

[a]KED: kidney essence deficiency.

[b]LFBU: liver fire bearing upward.

[c]PFSI: phlegm fire stagnation internally.

[d]WFAI: wind fire attacking internally.

[e]ODSS: Qi deficiency of the spleen and stomach.

Patient Similarity Scoring Based on Weighted Common Neighbor Algorithm

By transforming the TCM syndrome diagnostic problem into a prediction problem of linked patient nodes to TCM syndrome nodes, the similarity between 2 patients was calculated to obtain TCM syndrome similarity. For 2 patients, the higher the similarity, the greater the likelihood of having the same diagnostic result. This study measured the similarity using common features. In the knowledge graph, the higher the number of common neighbors to 2 patient nodes, the greater the likelihood of them belonging to the same community (linked to the same TCM syndrome node). The common neighbor graph of patients with different TCM syndromes is shown in Figure 7, where fewer common neighbors were observed. The common neighbor graph of patient 1 and patient 2 with the same TCM syndrome is shown in Figure 8, where more common neighbors were observed; however, different nodes had different importance. In TCM, the importance of pulse condition is greater than that of tinnitus duration while diagnosing tinnitus. The edge weight values of continuous tinnitus and thin pulse-to-kidney deficiency syndrome were 0.6991 and 1.1448, respectively, as shown in Figure 7; however, even for the same pulse condition, the importance varied for different TCM syndromes. In Figure 8, the edge weight values of thin pulse to QDSS and KED syndromes were 1.078 and 1.1447, respectively. Therefore, considering the edge weights of common neighbors to the patient nodes and calculating the score of common neighbors based on the edge weight values were essential when counting the number of common neighbors between patient nodes.

The similarity scoring function between patients x and y was defined by equation 6.

(6)

where $X = \{u_1, u_2, ..., u_m\}$ and $Y = \{v_1, v_2, ..., v_n\}$ represent the sets of neighboring nodes for patients x and y, respectively; $Path_{u,h,v} = (u, h, v)$ denotes the 2-hop path from node u to node v, where h represents the common neighbor of nodes u and v; $Path_{u,h} = (u, h)$ represents the path from node u to the common neighbor h; and $weight(path_{u,h})$ indicates the weight of the path.

When 2 paths with a hop count of 2 between the patient nodes existed, the weights of the paths were calculated to obtain a similarity score list for the patients. The list was then sorted in descending order, and the top 20 patient node syndromes with the highest scores were counted, which represented the most frequently occurring syndrome. Finally, the recommended syndrome was obtained.

$$S_n = G(f_{20}(\text{score}(X,Y))) \quad (7)$$

where G denotes a frequency-counting method in which X and Y represent sets of patient nodes. $f_{20}()$ was used to obtain the top 20 patient syndromes based on the scores.

Figure 7. Sketch map of common neighbors between different syndromes.

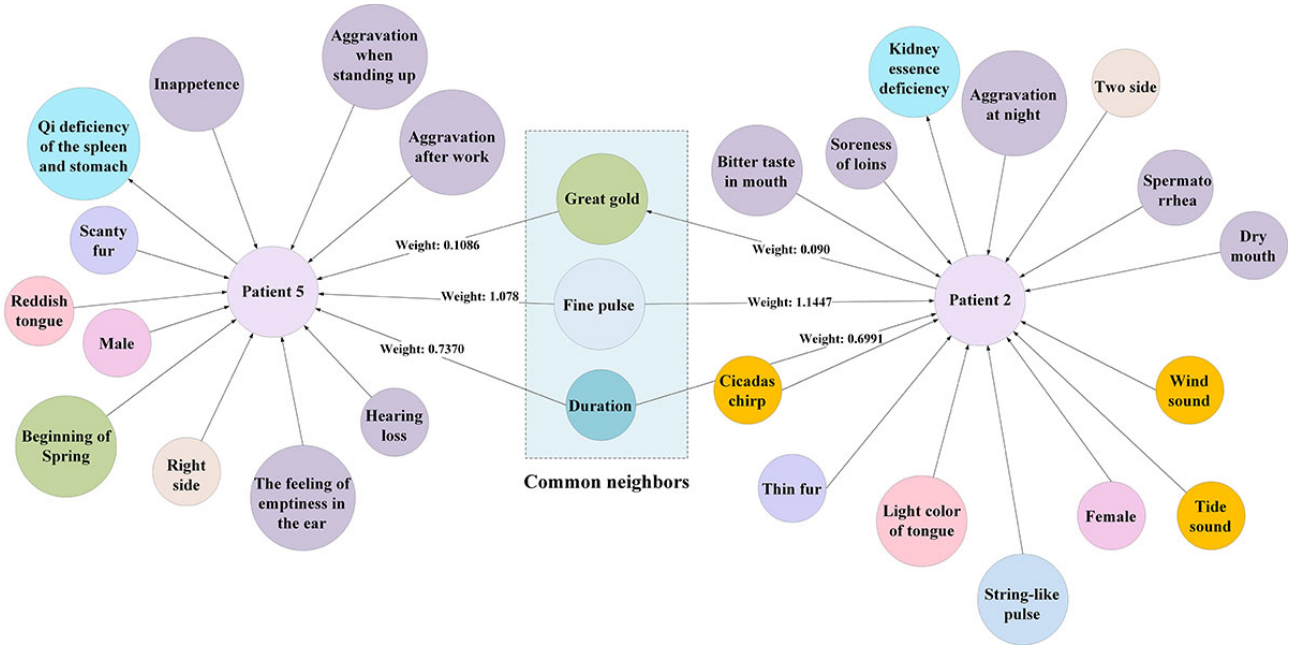
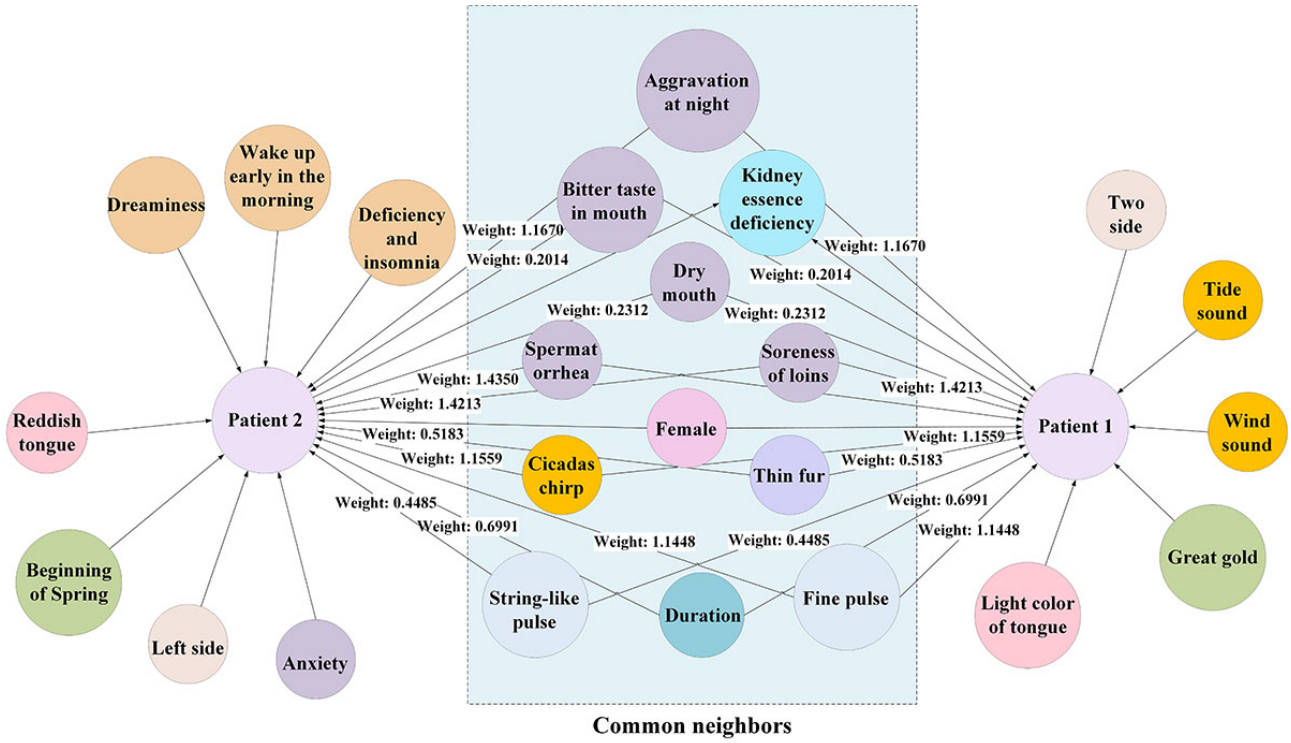


Figure 8. Sketch map of common neighbors between same syndromes.



Experimental Design

In total, 2 experiments were conducted to verify the effectiveness of the proposed method. The first experiment was performed to compare the proposed method with similar graph algorithms, while the second experiment was performed to compare the proposed method with other common explainable ML methods. The evaluation metrics of the algorithm are accuracy, precision, sensitivity, specificity, F_1-score, area under receiver operating characteristic curve (AUC), etc. To demonstrate the interpretability of our method, we selected a tinnitus case for result interpretation to showcase the inference process and interpretability of our method.

Results

Performance Verification

For a given knowledge graph, we extracted the patient nodes and their neighboring nodes to form a knowledge network. The node and edge sets in the knowledge network were divided into training and testing sets. The testing set did not contain syndrome entities. To reasonably divide the training and testing sets, we used a stratified sampling cross-validation method of

randomly dividing the network node and edge sets into 5 subsets: 1 subset as the testing set, and the other 4 subsets as the training set. The training set served as a known network, whereas the testing set was used to verify the syndrome prediction results and evaluate the accuracy of the syndrome prediction algorithm.

Evaluation Outcomes

Comparison With Similar Graph Algorithms

The proposed method was compared with similar graph algorithms such as CommonNeighbors and Adamic-Adar. CommonNeighbors is a common graph algorithm used to infer the potential relationships and proximity between 2 nodes [28]; however, the differences between common neighbors are not considered. Adamic-Adar is a typical algorithm for determining the closeness of 2 points by measuring the outdegree of common neighbors [29]. ResourceAllocation calculates the closeness between 2 nodes using a set of neighboring nodes near the target node [30]. We added common neighbor edge weights based on CommonNeighbors. Unlike Adamic-Adar and ResourceAllocation, our weight calculation method considered each syndrome, which had a higher adaptability to TCM diagnosis by the doctors. The experimental results are listed in Table 3; our method outperformed similar graph algorithms in diagnosing each syndrome.

Table 3. Experimental results of graph algorithm comparison.

Evaluation indicators and models	KED[a] (n=339)	LFBU[b] (n=307)	PFSI[c] (n=194)	QDSS[d] (n=270)	WFAI[e] (n=155)	Value, mean (SD)
Average accuracy						
Common neighbors	0.978	0.978	0.982	0.983	0.988	0.982 (0.004)
Adamic-Adar	0.979	0.979	0.978	0.983	0.989	0.982 (0.004)
Resource allocation	0.918	0.944	0.961	0.936	0.974	0.947 (0.019)
WeightedCommonNeighbors	0.990	0.994	0.995	0.992	0.998	0.994 (0.003)
Average precision						
Common neighbors	0.939	0.941	0.952	0.982	0.971	0.957 (0.017)
Adamic-Adar	0.940	0.949	0.932	0.981	0.971	0.955 (0.019)
Resource allocation	0.794	0.893	0.930	0.860	0.948	0.885 (0.055)
WeightedCommonNeighbors	0.970	0.987	0.993	0.986	1.000	0.987 (0.010)
Average sensitivity						
Common neighbors	0.981	0.971	0.922	0.943	0.929	0.949 (0.023)
Adamic-Adar	0.984	0.965	0.917	0.942	0.935	0.949 (0.023)
Resource allocation	0.933	0.877	0.801	0.840	0.837	0.857 (0.045)
WeightedCommonNeighbors	0.990	0.990	0.976	0.979	0.987	0.985 (0.006)
Average F_1-score						
Common neighbors	0.959	0.956	0.936	0.961	0.949	0.952 (0.009)
Adamic-Adar	0.961	0.957	0.924	0.961	0.952	0.951 (0.014)
Resource allocation	0.856	0.884	0.859	0.849	0.885	0.866 (0.015)
WeightedCommonNeighbors	0.980	0.989	0.984	0.982	0.994	0.986 (0.005)
Average specificity						
Common neighbors	0.978	0.980	0.993	0.995	0.996	0.988 (0.008)
Adamic-Adar	0.978	0.983	0.989	0.995	0.996	0.988 (0.007)
Resource allocation	0.914	0.966	0.990	0.963	0.994	0.965 (0.029)
WeightedCommonNeighbors	0.989	0.996	0.999	0.996	1.000	0.996 (0.004)
Average AUC[f]						
Common neighbors	0.979	0.976	0.958	0.969	0.963	0.969 (0.008)

Evaluation indicators and models	KED[a] (n=339)	LFBU[b] (n=307)	PFSI[c] (n=194)	QDSS[d] (n=270)	WFAI[e] (n=155)	Value, mean (SD)
Adamic-Adar	0.981	0.974	0.953	0.969	0.966	0.968 (0.009)
Resource allocation	0.923	0.922	0.895	0.901	0.915	0.911 (0.011)
WeightedCommonNeighbors	0.990	0.993	0.987	0.988	0.994	0.990 (0.003)

[a]KED: kidney essence deficiency.

[b]LFBU: liver fire bearing upward.

[c]PFSI: phlegm fire stagnation internally.

[d]QDSS: Qi deficiency of the spleen and stomach.

[e]WFAI: wind fire attacking internally.

[f]AUC: area under receiver operating characteristic curve.

Comparison With Other Interpretable ML Methods

The proposed method was compared with common ML classification algorithms including decision tree, random forest, naive Bayes, logistic regression, and k-nearest neighbors algorithms. The results are presented in Table 4. The graph algorithm based on WightedCommonNeighbor outperformed other models in the comprehensive diagnosis of each syndrome on the same data set but was lower than the random forest model in terms of the AUC metric. Although the random forest model had a certain degree of interpretability, the overall complexity of model interpretation increased when a large number of decision trees were included. The higher the number of decision trees in the random forest model, the greater the difficulty of interpreting the relationships and decision processes within the model. Compared to the random forest model, our proposed method had higher interpretability and was more readily accepted by doctors.

Table 4. Experimental results of machine learning classification algorithm comparison.

Evaluation indicators and models	KED[a]	LFBU[b]	PFSI[c]	QDSS[d]	WFAI[e]	Value, mean (SD)
Average accuracy						
WeightedCommonNeighbors	0.990	0.994	0.995	0.992	0.998	0.994 (0.003)
Decision tree	0.975	0.975	0.978	0.970	0.984	0.976 (0.005)
Random forest	0.987	0.982	0.985	0.987	0.994	0.987 (0.004)
Naive Bayes	0.979	0.976	0.979	0.981	0.991	0.981 (0.005)
Logistic regression	0.986	0.983	0.983	0.984	0.994	0.986 (0.004)
KNN[f]	0.986	0.980	0.982	0.986	0.994	0.985 (0.005)
Average precision						
WeightedCommonNeighbors	0.970	0.987	0.993	0.986	1.000	0.987 (0.010)
Decision tree	0.950	0.951	0.917	0.943	0.937	0.939 (0.012)
Random forest	0.974	0.950	0.970	0.982	0.963	0.968 (0.011)
Naive Bayes	0.971	0.923	0.953	0.956	0.980	0.957 (0.019)
Logistic regression	0.971	0.961	0.950	0.964	0.981	0.965 (0.010)
KNN	0.974	0.938	0.958	0.978	0.980	0.966 (0.016)
Average sensitivity						
WeightedCommonNeighbors	0.990	0.990	0.976	0.979	0.987	0.985 (0.006)
Decision tree	0.959	0.945	0.943	0.915	0.936	0.939 (0.014)
Random forest	0.976	0.977	0.933	0.956	0.987	0.966 (0.019)
Naive Bayes	0.953	0.981	0.912	0.956	0.948	0.950 (0.022)
Logistic regression	0.976	0.967	0.938	0.963	0.968	0.963 (0.013)
KNN	0.973	0.984	0.923	0.956	0.968	0.961 (0.021)
Average F_1-score						
WeightedCommonNeighbors	0.980	0.989	0.984	0.982	0.994	0.986 (0.005)
Decision tree	0.953	0.948	0.929	0.928	0.936	0.939 (0.010)
Random forest	0.975	0.963	0.950	0.968	0.975	0.966 (0.009)
Naive Bayes	0.961	0.951	0.932	0.955	0.964	0.953 (0.011)
Logistic regression	0.974	0.964	0.943	0.963	0.974	0.964 (0.011)
KNN	0.973	0.960	0.940	0.966	0.974	0.963 (0.012)
Average specificity						
WeightedCommonNeighbors	0.989	0.996	0.999	0.996	1.000	0.996 (0.004)
Decision tree	0.981	0.984	0.984	0.985	0.991	0.985 (0.003)
Random forest	0.990	0.983	0.994	0.995	0.995	0.992 (0.005)
Naive Bayes	0.989	0.974	0.992	0.988	0.997	0.988 (0.008)
Logistic regression	0.989	0.988	0.991	0.990	0.997	0.991 (0.003)
KNN	0.990	0.979	0.993	0.994	0.997	0.991 (0.006)
Average AUC[g]						
WeightedCommonNeighbors	0.990	0.993	0.987	0.988	0.994	0.990 (0.003)
Decision tree	0.970	0.964	0.964	0.950	0.963	0.962 (0.007)
Random forest	0.995	0.998	0.996	0.997	1.000	0.997 (0.002)
Naive Bayes	0.996	0.996	0.993	0.995	0.997	0.995 (0.001)
Logistic regression	0.997	0.997	0.994	0.995	0.997	0.996 (0.001)

Evaluation indicators and models	KED[a]	LFBU[b]	PFSI[c]	QDSS[d]	WFAI[e]	Value, mean (SD)
KNN	0.993	0.993	0.977	0.988	0.993	0.989 (0.006)

[a]KED: kidney essence deficiency.

[b]LFBU: liver fire bearing upward.

[c]PFSI: phlegm fire stagnation internally.

[d]QDSS: Qi deficiency of the spleen and stomach.

[e]WFAI: wind fire attacking internally.

[f]KNN: k-nearest neighbor.

[g]AUC: area under receiver operating characteristic curve.

Discussion

Principal Findings

The experimental results show that the accuracy, sensitivity, specificity, precision, F_1-score, and AUC of our proposed method all exceed 98% for 5 tinnitus subtypes. Compared to the traditional graph algorithm, our method comprehensively considers the number of neighboring nodes and the weight of edges for patient nodes. This method of calculating the strength of node connections and feature importance can more comprehensively measure the similarity between patient nodes. Further, by calculating the common neighbor score, the similarity between patient nodes can be quantitatively measured, providing a reliable quantitative indicator for the prediction problem of patient-to-syndrome node links. In addition, in the field of TCM, the impact of different features on diagnostic results may vary. This method considers the importance of features through edge weight values, making similarity calculations more realistic. By considering the edge weight values, the reasons for the formation of similarity between patient nodes and the importance of features can be explained, enhancing the interpretability of the model results. This method is not only applicable to the diagnosis of syndrome types in the field of TCM but can also be applied in other fields, especially in the similarity calculation problem that needs to consider feature importance and node correlation strength, which has universality.

In terms of interpretability, the proposed method integrated the knowledge of TCM differential diagnosis and clinical experience into a knowledge graph, which made the method more interpretable. To illustrate the explainability of our method, we randomly selected a patient from the patient records and used their medical information as input to the syndrome diagnosis algorithm, as shown in Figure 9. The patient information was input to the knowledge graph, where we searched for other patients who shared common neighbors with the selected patient. We calculated the common neighbor scores and returned the top k (k=20) patients with the highest scores. The results are summarized in Table 5. Based on the syndromes of the top k patients that were most similar to the target patient, we deduced that the predicted syndrome of the target patient was KED, which was consistent with the actual syndrome of the patient.

Figure 9. The inference process of patient syndrome patterns. KED: kidney essence deficiency.

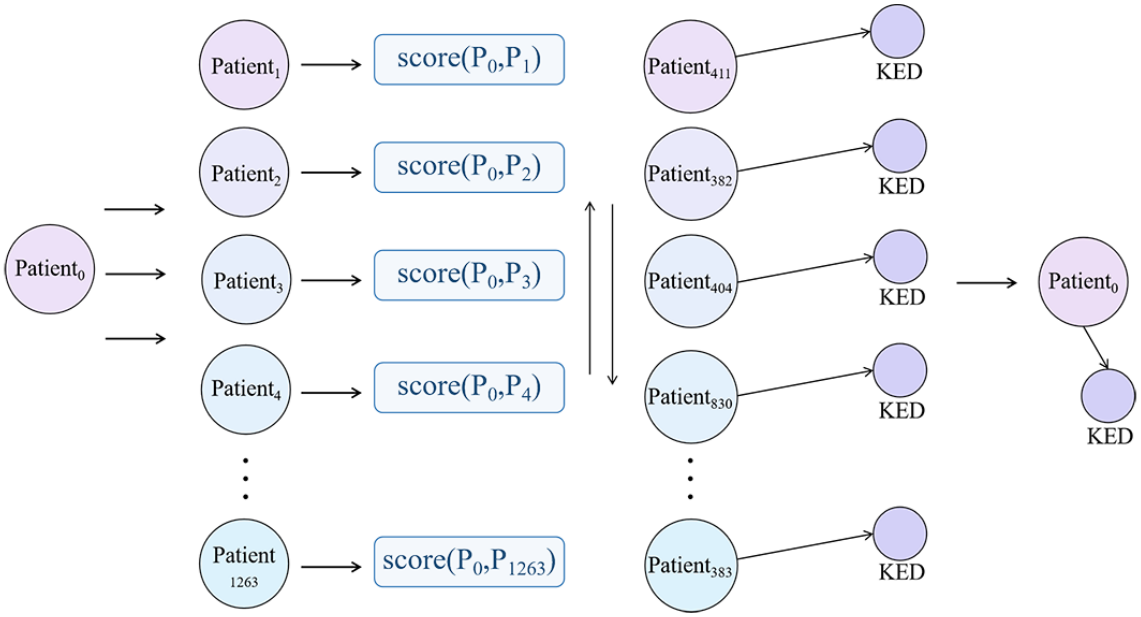


Table 5. Inference results of patient syndrome patterns.

Patient ID	Neighbors	Neighbors score
411	19	14.66
382	16	14.23
404	17	14.23
830	17	14.04
856	16	14.04
395	16	13.97
365	16	13.93
372	15	13.93
386	15	13.93
390	15	13.93
396	16	13.93
400	16	13.93
403	16	13.93
407	15	13.93
410	16	13.93
413	15	13.93
375	17	13.91
389	17	13.91
381	16	13.78
383	16	13.78

Limitations

The proposed method considered the weight of common neighbors and the importance of different symptoms for different syndrome types, but this makes similarity calculation more complex, requiring more computing resources and time. Meanwhile, the calculation of edge weight values requires relatively rich and accurate feature data. If the data quality is not high or features are missing, it will affect the accuracy of similarity calculation. However, compared to large-scale knowledge graphs, our research has a smaller sample size and requires continuous data collection to enrich the knowledge base.

From the experimental results, our method achieved good results in the diagnosis of WFAI, LFBU, PFSI, and QDSS. However, some deficiencies existed in the differential diagnosis of QDSS and KED syndrome types, which could create confusion between the two. The analysis of 3 patients who were misclassified with KED instead of QDSS revealed common entities between them and the top 5 most similar patients among their neighbors (Textbox 1). The common entities between patient 1 (ID 415) and the top 5 most similar patients among their neighbors, who were all patients with QDSS but were misclassified with KED, are listed in Textbox 1. The common entities included worsening conditions when standing up, empty feeling in the ears, left side, worsening condition after physical exertion, hypertension, red tongue, anxiety, thin pulse, hearing loss, continuous symptoms, female sex, and dizziness. Similarly, patient 2 (ID 601) and the top 5 most similar patients among their neighbors shared common entities including worsening condition when standing up, empty feeling in the ears, left side, worsening condition after physical exertion, thin and white coating on the tongue, red tongue, anxiety, thin pulse, and continuous symptoms. Patient 3 (ID 423) and the top 5 most similar patients among their neighbors shared common entities including worsening condition after physical exertion, worsening condition at night, left side, use of headphones, exercise, pale tongue, thin coating on the tongue, tinnitus, middle to low frequency, and intermittent symptoms. By comparing the common entities between the patients and their top 5 most similar neighbors, we found that entities such as worsening condition after physical exertion and left side had higher scores in the differential diagnosis of the 2 syndrome types. However, ML algorithms were prone to confusion in the differential diagnosis because both QDSS and KED could be present in patients with these symptoms.

Textbox 1. Misclassified patient entity.

Patient 1 (ID 415)

- Aggravation when standing up, ear emptiness, left side, aggravation after work, hypertension, tongue redness, anxiety, fine vein, hearing loss, duration, male, and dizziness.

Patient 2 (ID 601)

- Aggravation when standing up, ear emptiness, left side, aggravation after work, thin fur, white fur, tongue redness, anxiety, fine vein, and duration.

Patient 3 (ID 423)

- Aggravation after work, nighttime aggravation, left side, use headphones, exercise, tongue dullness, thin fur, cicada chirping, and interval.

Conclusions

Tinnitus is a complex ear disease that poses challenging issues in clinical diagnosis due to the lack of specific indicators and the reliance on patient complaints. In this study, we constructed a medical knowledge graph based on EMRs and authoritative knowledge of patients with tinnitus and proposed an explainable tinnitus-assisted diagnosis model. The experimental results showed that our proposed method not only performed better in diagnostic performance with a diagnostic accuracy of over 98% for all syndromes but also offered better interpretability compared to general ML algorithms owing to the natural interpretability of the knowledge graph. Thus, the effectiveness of the proposed method was demonstrated to assist Chinese medicine doctors in diagnosing tinnitus during clinical practice.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82074581).

Authors' Contributions

ZY and YG contributed to the conceptualization of this study and the funding acquisition. HZ and LW were responsible for data curation. ZY and ZK designed and implemented the algorithms and conducted the experiments. HZ, LW, ZK, TL, and ZW analyzed the experimental results. ZY wrote this paper with revision assistance from HZ and LW. YG reviewed and edited this paper. All authors have read and approved this paper.

Conflicts of Interest

None declared.

References

1. Jarach CM, Lugo A, Scala M, van den Brandt PA, Cederroth CR, Odone A, et al. Global prevalence and incidence of tinnitus: a systematic review and meta-analysis. JAMA Neurol 2022;79(9):888-900 [FREE Full text] [doi: 10.1001/jamaneurol.2022.2189] [Medline: 35939312]
2. Dawood F, Khan N, Bagwandin V. Management of adult patients with tinnitus: preparedness, perspectives and practices of audiologists. S Afr J Commun Disord 2019;66(1):e1-e10 [FREE Full text] [doi: 10.4102/sajcd.v66i1.621] [Medline: 31793315]
3. Ahmed A, Aqeel M, Akhtar T, Salim S, Ahmed B. Moderating role of stress, anxiety, and depression in the relationship between tinnitus and hearing loss among patients. Pak J Psychol Res 2019;34(4):753-772. [doi: 10.33824/pjpr.2019.34.4.41]
4. Neff P, Simões J, Psatha S, Nyamaa A, Boecking B, Rausch L, et al. The impact of tinnitus distress on cognition. Sci Rep 2021;11(1):2243. [doi: 10.1038/s41598-021-81728-0] [Medline: 33500489]
5. Piccirillo JF, Rodebaugh TL, Lenze EJ. Tinnitus. JAMA 2020;323(15):1497-1498. [doi: 10.1001/jama.2020.0697] [Medline: 32176246]
6. Özbey-Yücel Ü, Uçar A. The role of obesity, nutrition, and physical activity on tinnitus: a narrative review. Obesity Med 2023 Jun;40:100491. [doi: 10.1016/j.obmed.2023.100491]
7. Liu Z, Li Y, Yao L, Lucas M, Monaghan JJM, Zhang Y. Side-aware meta-learning for cross-dataset listener diagnosis with subjective tinnitus. IEEE Trans Neural Syst Rehabil Eng 2022;30:2352-2361. [doi: 10.1109/TNSRE.2022.3201158] [Medline: 35998167]
8. Sun ZR, Cai YX, Wang SJ, Wang C, Zheng Y, Chen Y, et al. Multi-view intact space learning for tinnitus classification in resting state EEG. Neural Process Lett 2019;49(2):611-624. [doi: 10.1007/s11063-018-9845-1]
9. Shoushtarian M, Alizadehsani R, Khosravi A, Acevedo N, McKay CM, Nahavandi S, et al. Objective measurement of tinnitus using functional near-infrared spectroscopy and machine learning. PLoS One 2020;15(11):e0241695 [FREE Full text] [doi: 10.1371/journal.pone.0241695] [Medline: 33206675]

10. Sanders PJ, Doborjeh ZG, Doborjeh MG, Kasabov NK, Searchfield GD. Prediction of acoustic residual inhibition of tinnitus using a brain-inspired spiking neural network model. Brain Sci 2021;11(1):52. [doi: 10.3390/brainsci11010052] [Medline: 33466500]
11. Manta O, Sarafidis M, Schlee W, Mazurek B, Matsopoulos GK, Koutsouris DD. Development of machine-learning models for tinnitus-related distress classification using wavelet-transformed auditory evoked potential signals and clinical data. J Clin Med 2023;12(11):3843 [FREE Full text] [doi: 10.3390/jcm12113843] [Medline: 37298037]
12. Allgaier J, Schlee W, Probst T, Pryss R. Prediction of tinnitus perception based on daily life mHealth data using country origin and season. J Clin Med 2022 Jul 22;11(15):4270 [FREE Full text] [doi: 10.3390/jcm11154270] [Medline: 35893370]
13. Rodrigo H, Beukes E, Andersson G, Manchaiah V. Exploratory data mining techniques (decision tree models) for examining the impact of internet-based cognitive behavioral therapy for tinnitus: machine learning approach. J Med Internet Res 2021;23(11):e28999 [FREE Full text] [doi: 10.2196/28999] [Medline: 34726612]
14. Liu Y, Niu H, Zhu J, Zhao P, Yin H, Ding H, et al. Morphological neuroimaging biomarkers for tinnitus: evidence obtained by applying machine learning. Neural Plast 2019;2019:1712342. [doi: 10.1155/2019/1712342] [Medline: 31915431]
15. Niemann U, Brueggemann P, Boecking B, Mazurek B, Spiliopoulou M. Development and internal validation of a depression severity prediction model for tinnitus patients based on questionnaire responses and socio-demographics. Sci Rep 2020;10(1):4664 [FREE Full text] [doi: 10.1038/s41598-020-61593-z] [Medline: 32170136]
16. Li Z, Fu Z, Li W, Fan H, Li S, Wang X, et al. Prediction of diabetic macular edema using knowledge graph. Diagnostics (Basel) 2023;13(11):1858 [FREE Full text] [doi: 10.3390/diagnostics13111858] [Medline: 37296709]
17. Zhou G, Kuang Z, Tan L, Xie X, Li J, Luo H. Clinical decision support system for hypertension medication based on knowledge graph. Comput Methods Programs Biomed 2022;227:107220. [doi: 10.1016/j.cmpb.2022.107220] [Medline: 36371975]
18. Lyu K, Tian Y, Shang Y, Zhou T, Yang Z, Liu Q, et al. Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy. J Biomed Inform 2023;139:104298 [FREE Full text] [doi: 10.1016/j.jbi.2023.104298] [Medline: 36731730]
19. Lin Z, Hong L, Cai X, Chen S, Shao Z, Huang Y, et al. Risk detection of clinical medication based on knowledge graph reasoning. CCF Trans Pervasive Comput Interact 2023;5(1):82-97. [doi: 10.1007/s42486-022-00114-5]
20. Yang YM, Li Y, Zhong X. Research on entity recognition and knowledge graph construction based on TCM medical records. J Artif Intell Pract 2021;47(1):1-15. [doi: 10.23977/jaip.2020.040105]
21. Xie Y, Hu L, Chen X. Auxiliary diagnosis based on the knowledge graph of TCM syndrome. Comput Mater Contin 2020;65:481-494. [doi: 10.32604/cmc.2020.010297]
22. Yang R, Ye Q, Cheng C, Zhang S, Lan Y, Zou J. Decision-making system for the diagnosis of syndrome based on traditional Chinese medicine knowledge graph. Evid Based Complement Alternat Med 2022;2022:8693937 [FREE Full text] [doi: 10.1155/2022/8693937] [Medline: 35186106]
23. Lan G, Hu M, Li Y, Zhang Y. Contrastive knowledge integrated graph neural networks for Chinese medical text classification. Eng Appl Artif Intell 2023;122:106057. [doi: 10.1016/j.engappai.2023.106057]
24. Liu D, Wei C, Xia S, YAN J. Construction and application of knowledge graph of Treatise on Febrile Diseases. Digital Chin Med 2022;5(4):394-405. [doi: 10.1016/j.dcmed.2022.12.006]
25. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 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); Jun 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
26. Liu P. Traditional Chinese Otorhinolaryngology. Beijing, China: China Traditional Chinese Medicine Press; 2021:90-94.
27. Wang DJ, Gan ZW. Traditional Chinese Otorhinolaryngology. Shanghai, China: Shanghai Scientific and Technical Publishers; 1985:26-28.
28. Freeman LC. A set of measures of centrality based on betweenness. Sociometry 1977;40(1):35-41. [doi: 10.2307/3033543]
29. Adamic LA, Adar E. Friends and neighbors on the web. Soc Networks 2003;25(3):211-230. [doi: 10.1016/s0378-8733(03)00009-1]
30. Mastrandrea R, Squartini T, Fagiolo G, Garlaschelli D. Enhanced reconstruction of weighted networks from strengths and degrees. New J Phys 2014;16(4):043022. [doi: 10.1088/1367-2630/16/4/043022]

Abbreviations

AI: artificial intelligence
AUC: area under receiver operating characteristic curve
EMR: electronic medical record
KED: kidney essence deficiency
LFBU: liver fire bearing upward
ML: machine learning
PFSI: phlegm fire stagnation internally

QDSS: Qi deficiency of the spleen and stomach
TCM: traditional Chinese medicine
WFAI: wind fire attacking internally

Edited by G Eysenbach, A Benis; submitted 05.03.24; peer-reviewed by L Wang, A Tomar, SN Mohanty; comments to author 20.04.24; revised version received 10.05.24; accepted 15.05.24; published 10.06.24.

Please cite as:
Yin Z, Kuang Z, Zhang H, Guo Y, Li T, Wu Z, Wang L
Explainable AI Method for Tinnitus Diagnosis via Neighbor-Augmented Knowledge Graph and Traditional Chinese Medicine: Development and Validation Study
JMIR Med Inform 2024;12:e57678
URL: https://medinform.jmir.org/2024/1/e57678
doi:10.2196/57678
PMID:38857077

Original Paper

Retrieval-Based Diagnostic Decision Support: Mixed Methods Study

Tassallah Abdullahi[1], MSc; Laura Mercurio[2], MD; Ritambhara Singh[1,3], PhD; Carsten Eickhoff[4], PhD

[1]Department of Computer Science, Brown University, Providence, RI, United States

[2]Departments of Pediatrics & Emergency Medicine, Alpert Medical School, Brown University, Providence, RI, United States

[3]Center for Computational Molecular Biology, Brown University, Providence, RI, United States

[4]School of Medicine, University of Tübingen, Tübingen, Germany

Corresponding Author:
Carsten Eickhoff, PhD
School of Medicine
University of Tübingen
Schaffhausenstr, 77
Tübingen, 72072
Germany
Phone: 49 7071 29 843
Email: carsten.eickhoff@uni-tuebingen.de

Abstract

Background: Diagnostic errors pose significant health risks and contribute to patient mortality. With the growing accessibility of electronic health records, machine learning models offer a promising avenue for enhancing diagnosis quality. Current research has primarily focused on a limited set of diseases with ample training data, neglecting diagnostic scenarios with limited data availability.

Objective: This study aims to develop an information retrieval (IR)–based framework that accommodates data sparsity to facilitate broader diagnostic decision support.

Methods: We introduced an IR-based diagnostic decision support framework called CliniqIR. It uses clinical text records, the Unified Medical Language System Metathesaurus, and 33 million PubMed abstracts to classify a broad spectrum of diagnoses independent of training data availability. CliniqIR is designed to be compatible with any IR framework. Therefore, we implemented it using both dense and sparse retrieval approaches. We compared CliniqIR's performance to that of pretrained clinical transformer models such as Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT) in supervised and zero-shot settings. Subsequently, we combined the strength of supervised fine-tuned ClinicalBERT and CliniqIR to build an ensemble framework that delivers state-of-the-art diagnostic predictions.

Results: On a complex diagnosis data set (DC3) without any training data, CliniqIR models returned the correct diagnosis within their top 3 predictions. On the Medical Information Mart for Intensive Care III data set, CliniqIR models surpassed ClinicalBERT in predicting diagnoses with <5 training samples by an average difference in mean reciprocal rank of 0.10. In a zero-shot setting where models received no disease-specific training, CliniqIR still outperformed the pretrained transformer models with a greater mean reciprocal rank of at least 0.10. Furthermore, in most conditions, our ensemble framework surpassed the performance of its individual components, demonstrating its enhanced ability to make precise diagnostic predictions.

Conclusions: Our experiments highlight the importance of IR in leveraging unstructured knowledge resources to identify infrequently encountered diagnoses. In addition, our ensemble framework benefits from combining the complementary strengths of the supervised and retrieval-based models to diagnose a broad spectrum of diseases.

(JMIR Med Inform 2024;12:e50209) doi:10.2196/50209

KEYWORDS

clinical decision support; rare diseases; ensemble learning; retrieval-augmented learning; machine learning; electronic health records; natural language processing; retrieval augmented generation; RAG; electronic health record; EHR; data sparsity; information retrieval

Introduction

Background

Identifying an accurate and timely cause for a patient's health problem represents a challenging and complex cognitive task. A clinician must consider a complex range of composite information sources, including the patient's medical history, current state, imaging, laboratory test results, and other clinical observations, to formulate an accurate diagnosis. Diagnostic errors are a leading cause of delayed treatment, potentially affecting millions of patients each year. Research suggests that these errors contribute to 6% to 17% of adverse events [1].

Studies [2,3] have shown that, rather than relying on a single physician for a final diagnosis, obtaining recommendations from multiple physicians increases diagnostic accuracy. To improve the diagnostic process while maintaining economic feasibility, different variants of automated assistants, also known as diagnostic decision support systems (DDSSs) and symptom checkers, have been introduced [4]. Early DDSSs [1,5] were driven by structured databases that maintain information about diseases and other medical information in a structured form. Although promising, these systems have yet to be highly successful for several reasons, including limited accessibility, poor flexibility, and scalability issues [6,7]. Hence, the traditional DDSS is gradually being replaced by machine learning and deep learning models.

Recent studies [8-13] highlight the importance of electronic health records for supervised machine learning algorithms in health care. These algorithms use the electronic health record of a patient as input to predict their diagnosis. However, supervised model development has been limited to a select number of diseases with higher prevalence and extensive documentation due to the availability of large amounts of labeled data. As a result, infrequently occurring diagnoses remain poorly studied. In real-world diagnostic scenarios, physicians are faced with the challenge of identifying the correct diagnosis from a plethora of possibilities. Therefore, a system that considers a broad range of diagnoses, including rare conditions, is desirable for improved diagnostic accuracy. However, recent studies [14,15] demonstrate that traditional supervised learning models are challenging to use in such scenarios due to their reliance on large, labeled data sets with many examples per diagnosis. However, most clinical cohorts exhibit imbalanced class distributions, characterized by a long-tailed pattern [15,16] in which certain diagnostic classes represent most training samples whereas others exhibit few or even 0 data points. In such scenarios, most traditional supervised models overfit the majority class, resulting in poor performance for the minority classes. As such, large labeled data sets may not be a straightforward solution for achieving an efficient supervised classifier that supports diverse diagnoses.

In response, researchers have leveraged a technique called transfer learning, which is a widely used method for building classifiers that enables generalization to classes with limited labeled data. A common transfer learning technique involves fine-tuning pretrained models—models trained on large and diverse data sets—on a smaller, domain-specific corpus to enhance model performance. However, the effectiveness of this approach still relies on the size of the data set available for fine-tuning. Zero-shot learning and few-shot learning [17,18] represent promising alternatives for fine-tuning large models with limited labeled data. In zero-shot learning, the model can classify samples from classes without labeled training data. Few-shot learning requires at least one labeled example per class to enable the model to make accurate predictions. Although some studies [19,20] have shown that pretrained language models have zero-shot and few-shot learning capabilities, their performance remains inferior to that of models trained on extensive labeled data. While zero-shot and few-shot approaches have demonstrated success in the vision domain [21,22], their application to language models remains an ongoing area of research.

Leveraging external knowledge resources can improve predictive performance, especially with a limited training sample size, as shown in previous work by Prakash et al [7] and Müller et al [6]. Classical information retrieval (IR) systems can use a vast collection of resources for various applications with low computational complexity and no need for labeled data. In the medical setting, studies [23-25] and competitions such as the text retrieval conference (TREC) clinical decision support track [26] have focused on developing and evaluating IR systems to support clinician decision-making. Typically, these IR systems have been applied to biomedical literature retrieval to aid in clinical decision support. However, these systems can also be adapted for other downstream clinical tasks. For example, Naik et al [27] trained a model to predict patient admission outcomes (ventilation need, mortality, and length of stay) by integrating relevant medical literature with patient notes, leaving an open question of how IR systems would fare in directly predicting the underlying diagnosis. Therefore, our study applied IR techniques to perform literature-guided diagnostic prediction.

Objectives

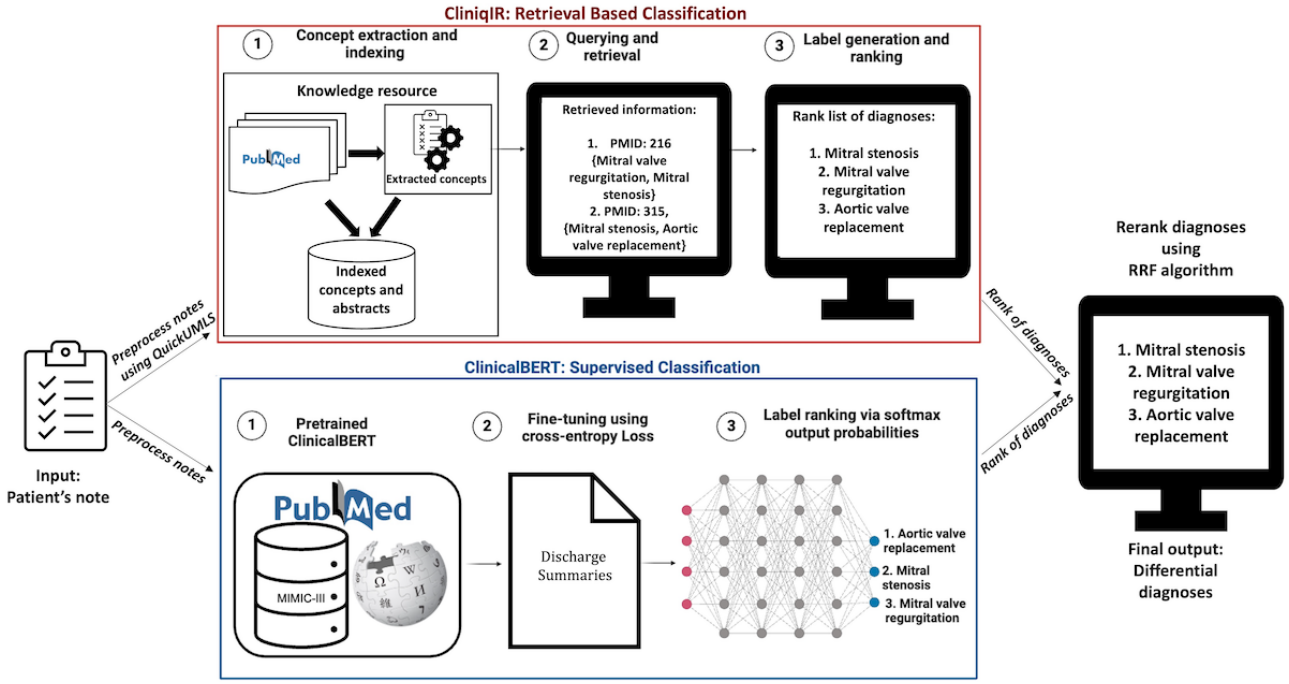
We introduce "CliniqIR," a novel clinical decision support algorithm that uses an IR system to match a patient's medical record to a specific diagnosis from a large pool of possible diagnoses. Our study aimed to improve the current state of predictive modeling and diagnostic decision support for a broad range of diagnoses regardless of their training data availability. By using clinical text records and external knowledge sources, including the Unified Medical Language System (UMLS) Metathesaurus [28] and PubMed abstracts [29], we demonstrated that "CliniqIR" successfully generalizes to less common diagnostic categories with heavily skewed data distributions. Our work also shows CliniqIR to be highly adaptable, allowing for easy integration with any IR system. This flexibility ensures the model's ability to adapt to available resources and work across various retrieval methods.

To assess CliniqIR's ability to predict diagnoses with no available training samples, we evaluated its performance on the DC3 data set [30]. We compared its performance to that of pretrained clinical models in a zero-shot setting, and our results showed that CliniqIR has the capability to recognize a broad spectrum of rare and complex diseases without relying on labeled training data. We also compared the performance of

CliniqIR with that of supervised fine-tuned pretrained biomedical large language models and found that supervised models have limitations when used on highly imbalanced data, especially for diagnoses with limited training samples. Then, we leveraged an ensemble strategy combining CliniqIR and a fine-tuned Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT) to make predictions for a wide range of diagnoses that include frequent and infrequent conditions, summarized in Figure 1.

Our study highlights the valuable synergy between retrieval-based systems and supervised learning models, showcasing how their combination can achieve state-of-the-art performance, particularly in data sets characterized by a long-tailed distribution. This finding holds significant promise and offers new avenues to address the challenges of imbalanced data in various domains.

Figure 1. CliniqIR and Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT), classify patient notes and generate ranked lists of potential diagnoses. The reciprocal rank fusion (RRF) ensemble reranks the lists from both models to provide clinicians with a more accurate final ranking of differential diagnoses to aid the diagnostic process. MIMIC-III: Medical Information Mart for Intensive Care III; PMID: PubMed ID.



Methods

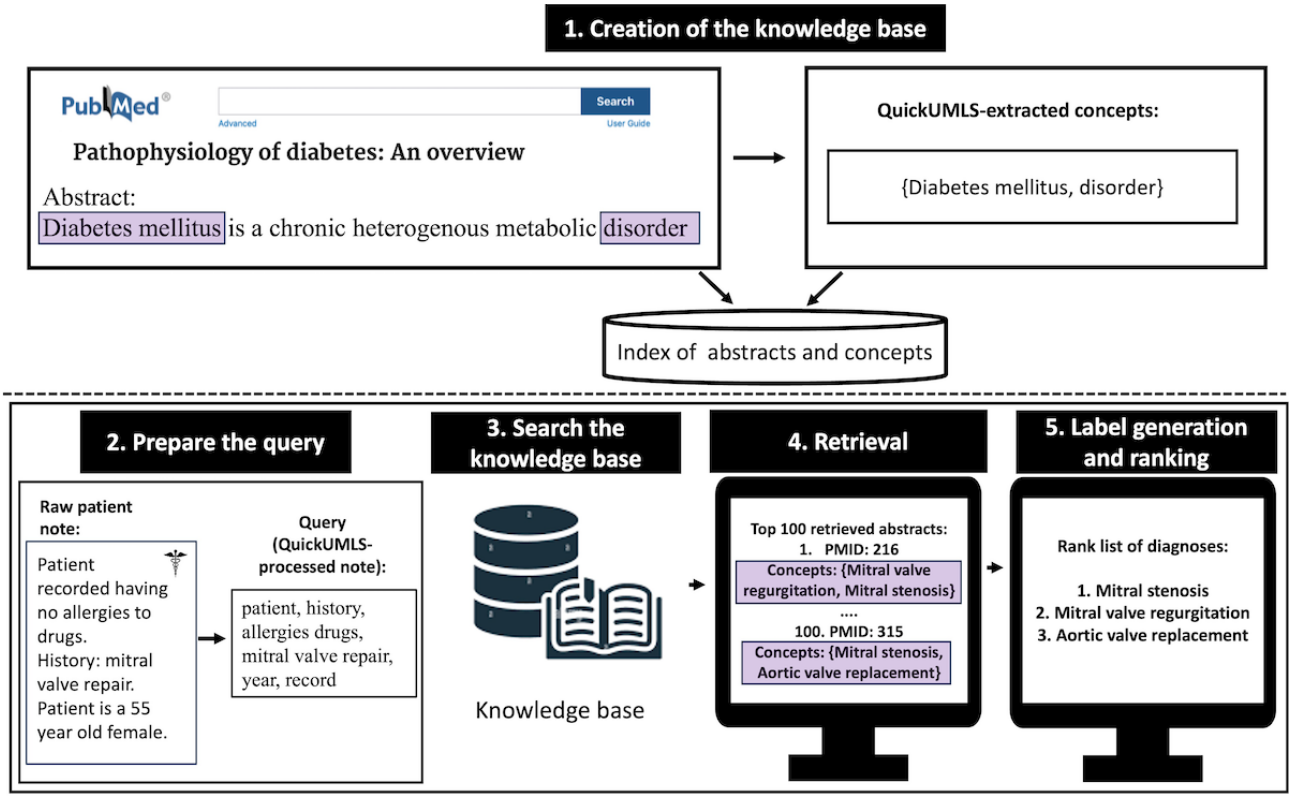
CliniqIR: The Retrieval-Based Model

Overview

We present CliniqIR, a novel literature-guided system that maps a patient's note to a specific diagnosis. By leveraging unlabeled external knowledge sources, CliniqIR uses an IR system to classify a wide range of diagnoses without relying on the availability of notes for each individual diagnosis (labeled training data). As a result, CliniqIR represents a valuable disease classification tool when labeled training data are limited or unavailable.

An overview of our method is shown in Figure 2. The backbone of CliniqIR is its knowledge base. Once the knowledge base is built, we can query the system to provide a list of probable diagnoses. In this study, a clinical narrative with a patient's medical history or summary was preprocessed and treated as a query. To make inferences given a patient's clinical note, as a preprocessing step, we first used QuickUMLS (Soldani and Goharian [31]) explained in the *Knowledge Extraction Using QuickUMLS* section, to extract medical keywords from the note to obtain a query. Next, we fed the query (preprocessed note) to the retrieval system, which returned a list of matching relevant PubMed abstracts alongside the medical conditions mentioned in each abstract. Afterward, we selected the top 100 items from the list and then computed the frequency of each concept across the list of abstracts. Finally, the model returned a list of concepts ranked according to their term frequency–inverse document frequency (TF-IDF) defined in equation 1. The list returned was similar to a medical differential diagnosis (a ranked list of possible diagnoses that could cause a patient's illness). The medical condition with the highest TF-IDF score was predicted as the most likely diagnosis. We provide a detailed description of the individual processing steps in the following sections.

Figure 2. Overview of CliniqIR, the retrieval-based clinical decision support system. PMID: PubMed ID.



Concept Extraction and Indexing

We extracted unique medical concepts (conditions) from each PubMed abstract using QuickUMLS, described in the *Knowledge Extraction Using QuickUMLS* section. Medical concepts included diseases, symptoms, or any information about a medical procedure. Subsequently, we built the knowledge base of the retrieval system by indexing each abstract and its corresponding article title, article ID, and a concept dictionary that contained all the unique concepts mentioned in that abstract. Indexing involves storing and organizing data to enable efficient IR at search time. Using the index of PubMed abstracts, the model inputs a patient's notes as a query and returns relevant information from the indexed abstracts as an output. Figure 2 provides visual details.

Querying and Retrieval

Once the index was built, we submitted queries to the retrieval system. The *Retrieval System Implementation* section provides more details. After we submitted a query, the system returned a list of abstracts and their corresponding attributes (dictionary of concepts, article title, and article ID number) ranked according to query relevance. For each query, we selected the top 100 abstracts because the top few documents are most likely to contain relevant query information.

Label Generation

After the querying operation, we focused on the extracted concepts of the top 100 abstracts. The previous retrieval phase can potentially return multiple abstracts that contain similar information in response to a given query, resulting in concept dictionaries of ≥2 abstracts containing similar concepts. Multiple occurrences could indicate the relevance of a concept across abstracts. To account for such duplication, we calculated each unique concept's recurrence, or term frequency (TF), across the list. The TF of a concept across a list of abstracts would be 1 if it appeared in only 1 abstract. If it appeared in 2 abstracts, its TF would become 2, and so on. Calculating the recurrence of concepts across the top-100 list resulted in a new list that contained medical concepts and their TFs. These medical concepts were regarded as labels and used for classification purposes. Thus, each unique concept became a potential diagnosis, and the TF of each concept is subsequently used for ranking purposes in equation 1. Textbox 1 describes the concepts the model returned (in no order of importance) after the retrieval stage given a set of queries processed using QuickUMLS. The list was filtered for a simple illustration. As mentioned previously, concepts are biomedical terms that include symptoms, signs, and diseases, among other things. On the other hand, a diagnosis could represent a disease, an injury, a neoplastic process, or a medical term describing a condition a patient is experiencing. Therefore, to account for a wide range of possible diagnoses, we kept all concepts in the label generation phase, and we considered a concept as a diagnosis when it matched the ground truth. Therefore, in this paper, we use *concepts* and *diagnoses* interchangeably.

Textbox 1. The output returned by the retrieval-based model (CliniqIR) given a query.

Query and concepts retrieved (labels)

- Abdominal pain, bloating, rectal bleeding, weight loss, anxiety, disruptive thoughts, and suicidality: "generalized anxiety disorder," "panniculitis," "chronic abdominal pain," "Burkitt's lymphoma," and "Whipple's disease"
- Chest pain, radiation to neck, dyslipidemia, lung crackles, bradycardia, and ST elevation: "acute myocardial infarction," "acute coronary syndrome," "coronary artery disease," "myocardial ischemia," "myopericarditis," and "myocardial infarctions"
- Night sweats, abdominal pain (pleuritic), nausea, loose stools, lymphadenopathy (inguinal), plaques, leucopenia, neutrophilia, and elevated (Angiotensin converting enzyme) ACE: "sarcoidosis," "lymphomas," "lymph node," "tuberculosis," "lupus erythematosus," "Rosai-Dorfman disease," and "Kikuchi-Fujimoto disease"

Ranking and Predictions

It is important to note that our model differs from traditional classification schemes. In our case, the observed mappings between patients' notes and ground-truth diagnoses are not provided for learning purposes. Therefore, a list of relevant diagnoses (a subset of the retrieved concepts) must be generated independently for each query. However, as the diagnosis list is not generated based on ground truth, it may contain information that is not relevant to the data set to be evaluated. For example, given a data set with 3 possible ground-truth diagnoses—*lymphoma*, *coronary artery disease*, and *gastroenteritis*—the model might return concepts such as *coronary artery disease*, *myocardial infarction*, and *chest pain* in the label retrieval phase for a query whose ground truth is *coronary artery disease*. To address this and ensure a fair comparison with other classification models, we filtered the retrieved concepts during the evaluation and only kept diagnoses that were part of the ground truth. Therefore, in the aforementioned example, we filtered out *myocardial infarction* and *chest pain*. Then, we assigned ranks to the remainder of the diagnoses in the list using the TF-IDF function shown in equation 1:

$$\text{TFIDF}(c,a,d) = \text{TF}(c,a).\text{IDF}(c,d) \quad \textbf{(1)}$$

Knowledge Resources: PubMed Abstracts

Over the years, research in predictive modeling for diagnostic decision support has witnessed enormous success in transfer learning, particularly where a model leverages an auxiliary data source (often a knowledge base) to perform several predictive tasks. Some studies [7,22,32] have used resources such as Wikipedia and PubMed [29] to create systems that perform classification tasks or retrieve useful articles with specific information. In contrast, most early DDSSs [1,33] were built on structured knowledge bases; however, most computable knowledge bases are not freely accessible.

Inspired by previous research, we used abstracts from PubMed articles as an unstructured collection of knowledge resources to guide the prediction of diagnoses for all our experiments. An abstract may contain information about a specific condition, its signs, or its symptoms. Some abstracts include medical case reports, whereas others may contain information about a medical device. To build a retrieval system grounded in reliable information, we leveraged the vast collection of abstracts in the PubMed database. PubMed, maintained by the National Library of Medicine, houses >33 million citations for biomedical literature, encompassing life science journals and books dating back to 1946. However, the number of abstracts available per condition varies considerably. Therefore, for our core experiments, we implemented a 100-abstract inclusion threshold for diagnoses (Multimedia Appendix 1).

Knowledge Extraction Using QuickUMLS

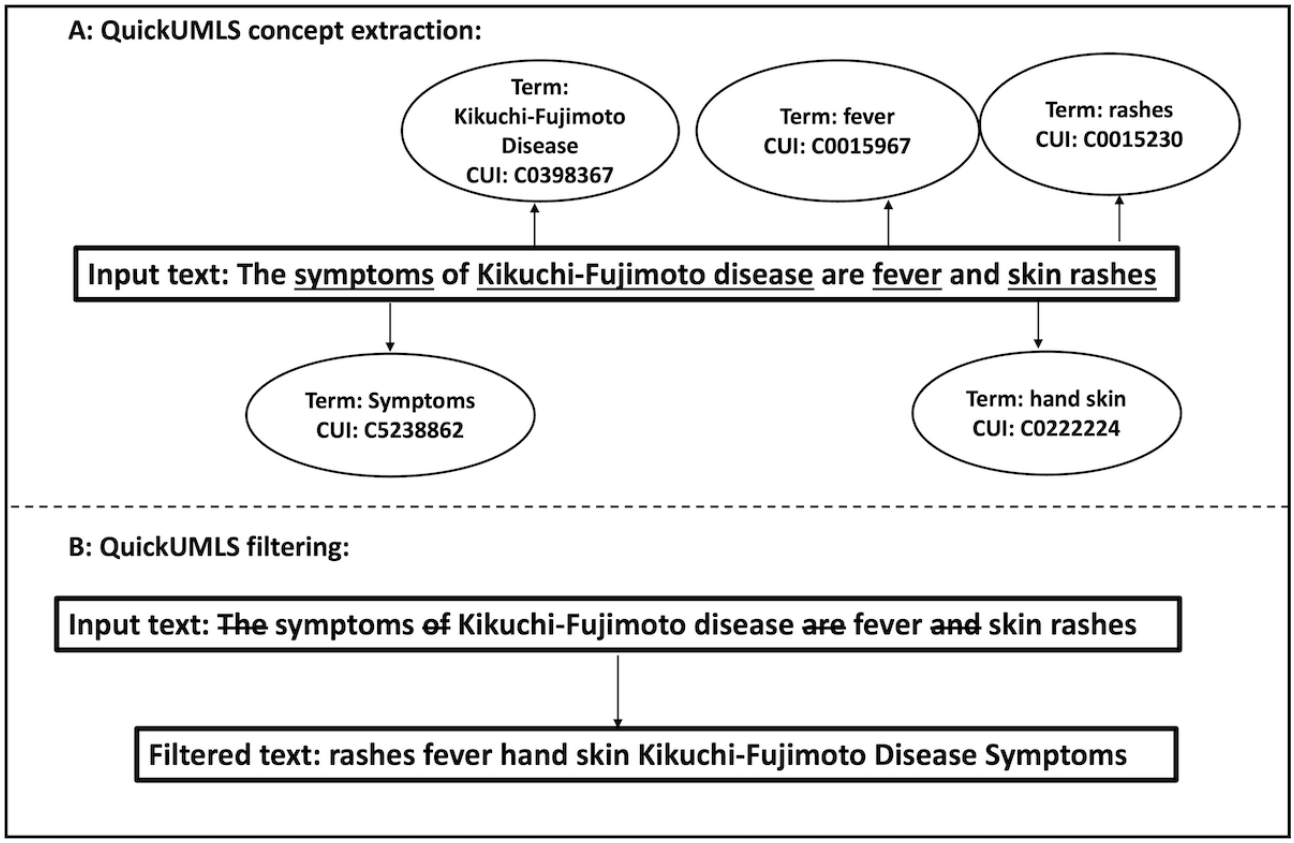
Overview

QuickUMLS [31] is an unsupervised medical concept extraction tool that detects mentions of medical entities such as diseases, symptoms, and other medical concepts from unstructured text. Given a document, QuickUMLS matches each possible token in the document against concepts in the UMLS [28]. In this study, we used QuickUMLS for 2 different purposes.

Extraction

We used QuickUMLS to extract unique biomedical concepts from each PubMed abstract. A concept can be any medical term, including a diagnosis or symptom. As shown in Figure 3, each biomedical term in a text is a concept with a corresponding unique alphanumeric identifier (concept ID) in the UMLS vocabulary. We kept all the concepts associated with each abstract in a dictionary.

Figure 3. Outputs from the QuickUMLS tool developed by Soldani and Goharian [33] showing: (A) a graph of extracted concepts and their concept unique identifier (CUI) for a specific input text; the underlined texts are considered important words, and their corresponding Unified Medical Language System terms and CUIs are returned, (B) a query processing pipeline. Each text marked with a strike-through is filtered out to obtain a query.



Filtering

We also used QuickUMLS as a data preprocessor to filter out noisy, uninformative, and nonclinical terms, such as stop words, from a patient's clinical note, resulting in a query that contained only medical terms. Citing the input text example in Figure 3, the outcome of the filtering operation was "rashes fever hand skin Kikuchi-Fujimoto disease symptoms." This filtering step is equivalent to keeping only the QuickUMLS-recognized medical terms and concepts.

Retrieval System Implementation

Overview

CliniqIR is designed to be highly adaptable to arbitrary IR systems. This flexibility ensures the model's ability to work across various retrieval methods, adapting to the resources available. In this study, we performed experiments on a sparse and a dense retriever.

Sparse Retriever

We built our knowledge base by indexing PubMed abstracts and their concepts using Apache Lucene (Apache Software Foundation [34]), which enables users to search this index with queries ranging from single words to sentences. The relevance of an abstract to a query is determined by a similarity score, with Lucene's default "BM25" [35] function estimating the best-matching abstract.

Dense Retriever

Unlike sparse retrievers, which represent queries as word frequencies, dense retrievers capture the semantic meaning and relationships within the text using dense embedding vectors. This allows for retrieval based on similarity, usually calculated through maximum inner-product search. To implement this approach, we leveraged the Medical Contrastive Pre-trained Transformers (MedCPT) [36], a state-of-the-art biomedical retrieval system in a zero-shot setting using its default parameters. Section S1 and Figure S1 in the Multimedia Appendix 1 provides details on the parameter settings for both retrieval systems.

Pretrained Transformer Models

In this study, we used 2 well-known methods, namely, supervised fine-tuning and zero-shot learning, to harness the benefits of transfer learning from 6 pretrained clinical and biomedical language models. The models we used are ClinicalBERT, PubMed Bidirectional Encoder Representations from Transformers (PubMedBERT), Scientific Bidirectional Encoder Representations from Transformers (SciBERT), Self-alignment Pretrained Bidirectional Encoder Representations from Transformers (SapBERT), cross-lingual knowledge-infused medical term embedding (CODER) and MedCPT. We describe them briefly in the follows:

ClinicalBERT Model

The ClinicalBERT [37] is an extension of Biomedical Bidirectional Encoder Representations from Transformers [38]

trained further on discharge summary notes from the Medical Information Mart for Intensive Care III (MIMIC-III) database [39]. It was designed to handle the complexity and nuances of clinical text.

PubMedBERT Model

PubMedBERT [40] was specifically designed to capture domain-specific knowledge present in biomedical literature. It was initialized from Bidirectional Encoder Representations From Transformers (BERT) and trained further on the collection of PubMed abstracts.

SciBERT Model

SciBERT [41] is a BERT-based language model pretrained on 1.14 million full-text papers from Semantic Scholar. The corpus domain cuts across the field of computer science and the biomedical space.

SapBERT Model

SapBERT [42] is also a BERT-based model initialized from PubMedBERT. SapBERT was further pretrained on UMLS [28], which consists of a wide range of biomedical ontologies for >4 million concepts.

CODER Model

CODER [43] is another BERT-based model formulated to generate biomedical embeddings. It was also initialized from PubMedBERT. CODER was further pretrained using the concepts from the UMLS [28] and optimized to increase the embedding similarities between terms with the same concept unique identifier.

MedCPT Model

MedCPT [36] is a contrastive pretrained PubMedBERT-based model also formulated to generate biomedical text embeddings for multiple tasks.

Supervised Fine-Tuning

Given a set of patients' notes (hereinafter also referred to as *notes*) as inputs and their corresponding diagnoses as outputs, we fine-tuned the pretrained models in a supervised fashion to classify diagnoses by feeding in a series of notes and their corresponding ground-truth diagnoses. Each note was a textual document describing a patient's health condition and medical history. The ground-truth diagnosis of a note was the corresponding health condition of the patient. Multimedia Appendix 1 provides details of the models' parameter settings. After fine-tuning, given a test set of notes, a model assigned probabilities to each ground-truth diagnosis for each note. The diagnosis with the highest probability corresponded to the model's most confident prediction. We assigned ranks to each diagnosis in the order of their decreasing probability score for all our predictions. These ranks were further used to compute the mean reciprocal rank (MRR) for model evaluation (refer to the *Evaluation Metrics* section for details). We justify the use of ranking output probabilities across classes to compute the MRR because the probabilities generated by the classifier represent the classifier's confidence in predicting each incidence. Supervised fine-tuning requires diagnosis-specific training data (availability of historic patient notes for each diagnosis) to deliver state-of-the-art performance. Unfortunately, labeled data are expensive to generate. This requirement makes it impractical to use a supervised fine-tuned model to diagnose those diseases without (many) notes for training. Hence, we used this method to make predictions only when training data were available.

Zero-Shot Learning

Given our focus on predicting diagnoses with few or 0 training samples, we included zero-shot learning methods as baselines. Leveraging the high quality of the aforementioned pretrained transformer embeddings, we adopted a zero-shot strategy by classifying patients' notes based on their semantic similarity to potential diagnoses. This can be achieved by using pretrained models as biencoders [18,44]. Using this approach, we accounted for the diagnosis classes (classes without training samples) that the supervised fine-tuned models could not handle.

Given a patient's note (our query) and the list of candidate diagnoses as labels, we used different variants of BERT as biencoders to encode queries and the full names of all ground-truth diagnoses to produce their respective representation vectors separately. Next, we computed their cosine similarity score and ranked each diagnosis for each query according to this score. The diagnosis with the highest cosine similarity became the model's most confident diagnostic prediction (refer to Multimedia Appendix 1 for more details).

Model Ensemble: Reciprocal Rank Fusion

The label retrieval process allowed the CliniqIR (retrieval-based model) to diagnose unseen conditions regardless of training data availability. This property is beneficial for diagnoses with little or no training data. On the other hand, a supervised fine-tuned model can draw much deeper insights from available historical case data. We adopted an ensemble strategy to combine the advantages of both paradigms.

In IR and general machine learning, ensemble strategies combine results from multiple models to produce a single joint output. Ideally, the ensemble model should produce a new output whose performance is superior to that of the individual constituent models. Several studies [32,45,46] have shown that high-performance gains can be achieved through model ensembling. One of the simplest ways to build such a model is to focus on applying a reranking heuristic to the ranks of each item in a model's output list. Hence, we collected the ranked list of diagnoses from a CliniqIR model and that of the best-performing supervised fine-tuned model, ClinicalBERT, and combined the 2 lists. We then applied a modified version of the reciprocal rank fusion (RRF) [45] algorithm using equation 2 to merge their results and produce a single, final output list. Given a set *C*of concepts(diagnoses) to be ranked and a set of rankings *R*for all concepts obtained from each ensemble member (CliniqIR and ClinicalBERT), we computed the RRF score for each concept ($c \in C$) as follows:

(2)

In the aforementioned equation, "$r \in R$" is the rank of concept *c*according to an ensemble member. We summed up the individual ranks of a concept from each ensemble member "$r(c)$"

with *k* and computed the inverse. Previous work by Cormack et al [45] reported that setting *k*to 60 was the near-optimal choice for most of their experiments. Hence, we set *k* to 60 for all experiments. When concepts (diagnoses) had more than one training sample, we selected their individual ranks *r*from each ensemble member to compute the RRF score; otherwise, we selected ranks from the CliniqIR model. We used the RRF algorithm due to some key advantages: (1) it is a simple unsupervised method that eliminates the need for training samples, and (2) it effectively combines the results from various models without reliance on a weighting or voting mechanism.

Experimental Setup

Data Sources

DC3 Data Set

The DC3 data set [30] was designed specifically for the evaluation of diagnostic support systems. The data set comprises 30 rare and difficult-to-diagnose cases compiled and solved by clinical experts in the *New England Journal of Medicine* Case Challenges. This data set lacks large, labeled training data, but it covers a wide range of diagnostic cases for various specialties. Therefore, we used this data set to determine the applicability of CliniqIR for diagnostic inference when the underlying patient condition is rare. Each case is a patient's note and its corresponding true diagnosis written as free text. We mapped the true diagnoses to their UMLS concept IDs to produce test labels for evaluation consistency. When we did not find an exact matching term for a diagnosis, we considered the closest match returned by the UMLS browser. During the preprocessing step, we found that some cases in the DC3 data set had multiple terms representing a ground-truth diagnosis, making it difficult to find a single UMLS concept ID for such cases. To ensure an accurate mapping with the UMLS concept IDs, we split such cases into separate terms. For example, the case "Acute and chronic cholecystitis and extensive cholelithiasis with transmural gallbladder inflammation" was split into 2 separate terms: "Acute and chronic cholecystitis" and "Extensive cholelithiasis with transmural gallbladder inflammation." Then, we mapped each case to its corresponding UMLS concept ID. Next, we computed the document frequency of all the true diagnoses (now represented as concepts) across all PubMed abstracts. In these cases, either of the concepts could be considered as the ground truth. As the data did not contain sufficient notes to train a model, we formulated this task as a zero-shot multiclassification problem. Specifically, we expected a model to predict the underlying condition given a patient's note without labeled training data.

MIMIC-III Data Set

The MIMIC-III [39] is a freely accessible medical database that contains information on >50,000 intensive care unit patients. The data include laboratory events, vital sign measurements, clinical observations, notes, and diagnoses structured as *ICD-9-CM* (*International Classification of Diseases, Ninth Revision, Clinical Modification*), codes. We worked with the discharge notes for all experiments because they document a free-text synopsis of a patient's hospital stay from admission to discharge. In MIMIC-III, each discharge note is mapped to multiple diagnoses ranked according to priority. We considered the highest-priority diagnosis to be the admission's ground-truth diagnosis (and prediction target). We excluded admissions primarily for birth and pregnancy as they did not represent a primary pathological diagnosis. After preprocessing, the discharge notes contained 2634 unique *ICD-9-CM* diagnoses. We mapped these *ICD-9-CM* diagnoses to their corresponding UMLS concept IDs to calculate their TF across the knowledge resource (PubMed abstracts). The resulting unique diagnoses were associated with notes ranging from thousands of occurrences of frequent conditions, such as coronary atherosclerosis and aortic valve disorders, to rare ones, such as Evans syndrome and ehrlichiosis, with just a single instance forming a long-tailed distribution. A total of 902 diagnoses fell into the singleton category. One discharge note representing a specific diagnosis is insufficient to train and test a model. Thus, we reserved all diagnoses with only 1 available note for model testing. For diagnoses with <5 note samples, we reserved 1 sample for testing, and the rest were included in model training. We split the remainder of the data set (instances of diagnoses with ≥5 associated notes) into training, validation, and testing sets in the ratio 70:15:15; this split resulted in the training set containing notes representing 1732 unique diagnoses and the test set containing notes representing a total of 2634 unique diagnoses (refer to Multimedia Appendix 1 for more details). For models that did not require training (eg, the retrieval model), we used the validation and training sets for hyperparameter tuning purposes and the test set for final model evaluations.

Baselines

Previous studies and competitions, such as the TREC clinical decision support track, have emphasized the development and evaluation of IR systems to aid clinical decision-making. While these systems are commonly used for evidence-based literature searches, our study explored their adaptation for direct literature-guided diagnosis prediction. Although a direct comparison to the systems in the TREC clinical decision support track was not possible, insights gained from these competitions informed the engineering of our retrieval system. To evaluate our model, we used 2 transfer learning techniques—supervised fine-tuning and zero-shot classification methods (refer to the *Pretrained Transformer Models* section)—because of their performance in scenarios where labeled data are limited or unavailable. In addition, some studies [47-49] have shown pretrained language models to attain superior performance to that of count vector–based models and traditional supervised methods in various medical tasks. We used "Z" to identify when models were used in a zero-shot classification setting, an "S" for supervised fine-tuning, and "CliniqIR" when models were used in a retrieval setting. Table 1 provides details.

Table 1. Overview of the experiments conducted using the different models and their task description.

Experiment	Models used	Task description
Retrieval-based experiments (CliniqIR)	BM25[a] and MedCPT[b]	Models retrieved relevant abstracts to inform diagnostic predictions.
Zero-shot experiments (Z)	ClinicalBERT[c], PubMedBERT[d], CODER[e], SapBERT[f], and MedCPT	Models classified diseases in a zero-shot setting without previous task-specific training.
Supervised experiments (S)	ClinicalBERT	Models were fine-tuned using labeled data for enhanced disease prediction accuracy.

[a]BM25: Best Match 25.

[b]MedCPT: Medical Contrastive Pretrained Transformers.

[c]ClinicalBERT: Clinical Bidirectional Encoder Representations from Transformers.

[d]PubMedBERT: PubMed Bidirectional Encoder Representations from Transformers.

[e]CODER: cross-lingual knowledge-infused medical term embedding.

[f]SapBERT: Self-alignment Pretrained Bidirectional Encoder Representations from Transformers.

Evaluation Metrics

In our experiments, each model returned a ranked list of diagnoses analogous to a ranked list of differential diagnoses formulated by a medical expert. Given a query and a list of ranked items produced by a model, a simple classification accuracy metric tracks whether the model made the correct prediction at the top of the list. Instead, we used the MRR [50] because it told us *where* the true diagnosis was placed in the list in equation 3. If a model returned the reference diagnosis at rank 1 (ie, at the top of the list), the reciprocal rank (RR) was 1; if the most appropriate item was at rank 2, then the RR was 0.5. The RR decreases as the relevant item moves farther down the list. We calculated the MRR by computing the average RR across admissions. An MRR of 1 meant that the model returned the correct diagnosis at the top of its list for every patient, and an MRR of 0 implied that the model never produced a correct diagnosis. Mathematically, the MRR can be represented as follows:

(3)

where $|Q|$ denotes the total number of queries and denotes the rank of the correct diagnosis. We also calculated the mean average precision (MAP) to evaluate the balance between precision and recall of the retrieval systems (for details, refer to Multimedia Appendix 1).

Ethical Considerations

No ethics approval was pursued for this research, given that the data were publicly accessible and deidentified.This aligns with the guidelines outlined by the US Department of Health and Human Services, Office for Human Research Protections, §46.101 (b)(4) [51].

Results

CliniqIR Models Retrieved Useful Literature and Meaningful Concepts

Table 2 showcases qualitative results for 3 selected queries, displaying the top 3 documents retrieved by the CliniqIR model. Notably, the retrieved articles and their corresponding concepts demonstrated clear relevance to the ground-truth diagnoses of the respective queries.

Table 2. Qualitative overview of the top documents and concepts retrieved for 3 selected queries along with their respective correct concepts. This table illustrates the types of results our system generates for each query, showing the alignment with the ground-truth concepts.

Ground-truth diagnosis, and top 3 documents	Retrieved concepts
1. Viral pneumonia	
Relevant article 1—PMID[a] 15336585: Cases from the Osler Medical Service at Johns Hopkins University. Diagnosis: P. carinii pneumonia and primary pulmonary sporotrichosis	{“C1956415”: [“paroxysmal nocturnal dyspnea”], “C0239295”: [“esophageal candidiasis”], “C0236053”: [“mucosal ulcers”], “C1535939”: [“Pneumocystis”], “C0031256”: [“petechiae”], “C0006849”: [“thrush”], “C0011168”: [“dysphagia”]}
Relevant article 2—PMID 32788269: A 16-Year-Old Boy with Cough and Fever in the Era of COVID-19	{“C0746102”: [“chronic lung disease”], “C0004096”: [“asthma”], “C0009443”: [“cold”], “C0206750”: [“Coronavirus”], “C0018609”: [“h disease”]}
Relevant article 3—PMID 30225154: Meningococcal Pneumonia in a Young Healthy Male	{“C3714636”: [“pneumonias”], C1535950”: [“GI inflammation”]}
2. Hypoparathyroidism	
Relevant article 1—PMID 34765380: A Challenging Case of Persisting Hypokalemia Secondary to Gitelman Syndrome	{“C0220983”: [“metabolic alkalosis”], “C0151723”: [“hypomagnesemia”], “C0020599”: [“hypocalciuria”], “C0014335”: [“enteritis”], “C0012634”: [“Diagnosis”], “C0235394”: [“wasting”], “C0271728”: [“Hyperreninemic hyperaldosteronism”], “C0268450”: [“gitelman syndrome”], “C3552462”: [“Tubulopathy”]}
Relevant article 2—PMID 27190662: Suppression of Parathyroid Hormone in a Patient with Severe Magnesium Depletion	{“C0151723”: [“hypomagnesemia”], “C0030554”: [“paresthesias”], “C0020598”: [“hypocalcemias”], “C0020626”: [“Low parathyroid hormone”], “C0030517”: [“Parathyroid”], “C0033806”: [“pseudo hypoparathyroidism”]}
Relevant article 3—PMID 28163524: Afebrile Seizures as Initial Symptom of Hypocalcemia Secondary to Hypoparathyroidism	{“C0020626”: [“Hypoparathyroidism”], “C0012236”: [“DiGeorge syndrome”], “C0863106”: [“afebrile seizures”], “C0030353”: [“papilledema”], “C0020598”: [“Hypocalcemias”], “C0012634”: [“Diagnosis”], “C0042870”: [“Vitamin D deficiency”]}
3. Intracerebral hemorrhage	
Relevant article 1—PMID 9125737: A 36-year-old woman with acute onset left hemiplegia and anosognosia	{“C0020564”: [“enlargement”], “C0019080”: [“hemorrhage”]}
Relevant article 2—PMID 25830084: Multiple extra-ischemic hemorrhages following intravenous thrombolysis in a patient with Trousseau syndrome: case study.	{“C2937358”: [“Intracerebral hemorrhage”], “C0151699”: [“intracranial hemorrhage”], “C0019080”: [“hemorrhages”], “C0020564”: [“enlargement”], “C0021308”: [“infarct”], “C0022116”: [“ischemia”]}
Relevant article 3—PMID 1434057: A case of recurrent cerebral hemorrhage considered to be cerebral amyloid angiopathy by cerebrospinal fluid examination.	{“C0472376”: [“thalamic hemorrhage”], “C2937358”: [“cerebral hemorrhage”], “C0019080”: [“bleeding”], “C0023182”: [“cerebrospinal fluid leak”]}

[a]PMID: PubMed ID.

CliniqIR Models Yielded State-of-the-Art Performance for Rare and Complex Diagnoses

We examined the retrieval-based models’ (CliniqIR) performance on the DC3 data set to show their applicability for rare and complex diagnostic cases. The absence of training data for this data set implied that supervised learning would not be applicable and the models could only make predictions in an unsupervised or zero-shot setting. Hence, on this data set, we compared the CliniqIR models’ performance to that of pretrained transformers in a zero-shot setting. In contrast to the CliniqIR model, which creates its own set of labels, we supplied the pretrained transformers with a range of potential diagnoses for each query to enable zero-shot predictions. This gave the models a significant advantage over their use in a real-world setting, where such information would not be readily available. Table 3 shows the MRR of the chosen models on the DC3 data set. Even with the supporting assumption that the range of possible diagnoses was known to the pretrained models, the CliniqIR models outperformed them with an MRR of 0.35 and 0.32 for CliniqIR_BM25 and CliniqIR_MedCPT, respectively. This means that, on average, CliniqIR_BM25 and CliniqIR_MedCPT were more likely to return the correct diagnosis within the top 3 predictions for a case.

The MRR scores of the pretrained zero-shot methods were similar to one another but markedly lower; the scores were 0.15, 0.22, 0.25, 0.25, 0.24, and 0.18 for ClinicalBERT, PubMedBERT, SciBERT, CODER, SapBERT, and MedCPT, respectively. Our results show that the CliniqIR models are capable of making useful predictions in the case of rare and complex diagnoses with limited or no training data availability.

Table 3. Performance evaluation of the models on the DC3 data sets across all case. The retrieval-based models, denoted using "CliniqIR" gave the best overall performance compared to the zero-shot models, denoted using "Z."

Model used	Mean reciprocal rank
ClinicalBERT[a] (Z)	0.15
PubMedMERT[b] (Z)	0.22
SciBERT[c] (Z)	0.25
CODER[d] (Z)	0.25
SapBERT[e] (Z)	0.24
CliniqIR_BM25	*0.35* [f]
MedCPT[g] (Z)	0.18
CliniqIR_MedCPT	*0.32*

[a]ClinicalBERT: Clinical Bidirectional Encoder Representations from Transformers.

[b]PubMedBERT: PubMed Bidirectional Encoder Representations from Transformers.

[c]SciBERT: Scientific Bidirectional Encoder Representations from Transformers.

[d]CODER: cross-lingual knowledge-infused medical term embedding.

[e]SapBERT: Self-alignment Pretrained Bidirectional Encoder Representations from Transformers.

[f]Highest mean reciprocal rank is italicized.

[g]MedCPT: Medical Contrastive Pre-trained Transformers.

Performance on MIMIC-III

Supervised Prediction Models Failed at Making Rare Diagnoses

When training data are available, supervised models are preferred. Thus, we investigated the effectiveness of a supervised learning approach for a highly imbalanced data set such as MIMIC-III. We fine-tuned the pretrained models to predict diagnoses using available clinical notes. Diagnoses were categorized based on the frequency of associated notes to show how training data availability affects a supervised model's predictive capacity. A total of 902 diagnoses had no training data (only 1 note representative in MIMIC-III), whereas 1732 had at least one training sample (≥2 note representatives in MIMIC-III). Predictions were made only for the 1732 diagnoses, excluding those without training data. We introduced sample weights in the loss function to handle the enormous data imbalance. This approach weighs the loss computed for samples differently depending on their class training size. Our results in Figure S2 in Multimedia Appendix 1 show that ClinicalBERT performed best among all pretrained models. Hence, we used ClinicalBERT as our supervised baseline for the remainder of our experiments.

After training ClincalBERT, we tested it on different clinical note frequency–based categories (Table 4). In Table 4, we observed that the MRR score of the ClinicalBERT model was higher for diagnosis categories with many training examples (>10 notes). In addition, in the data set category with 1 to 10 notes available per diagnosis (1<notes≤10), ClinicalBERT obtained a low MRR score of 0.07. An MRR of 0.08 indicates that, on average, ClinicalBERT returned the correct diagnosis for a case among its top 13 predictions for these diagnoses. While the model could not perform predictions for 902 diagnoses due to the lack of training data, the drastic decline in ClinicalBERT's performance also indicates that the model is not suitable for making predictions for diagnoses with <10 clinical notes available for training. We also noticed a decline in performance for diagnoses with training samples between 500 and 750. This was likely due to many diagnoses having similar symptoms and manifestations. Therefore, the supervised learning approaches struggle to find a fine delineation of boundaries between similar classes without sufficient training data.

Table 4. Performance of the best-performing fine-tuned supervised model, Clinical Bidirectional Encoder Representations from Transformers on the Medical Information Mart for Intensive Care III data set. We categorized the results by the frequency of training note representation per diagnosis.

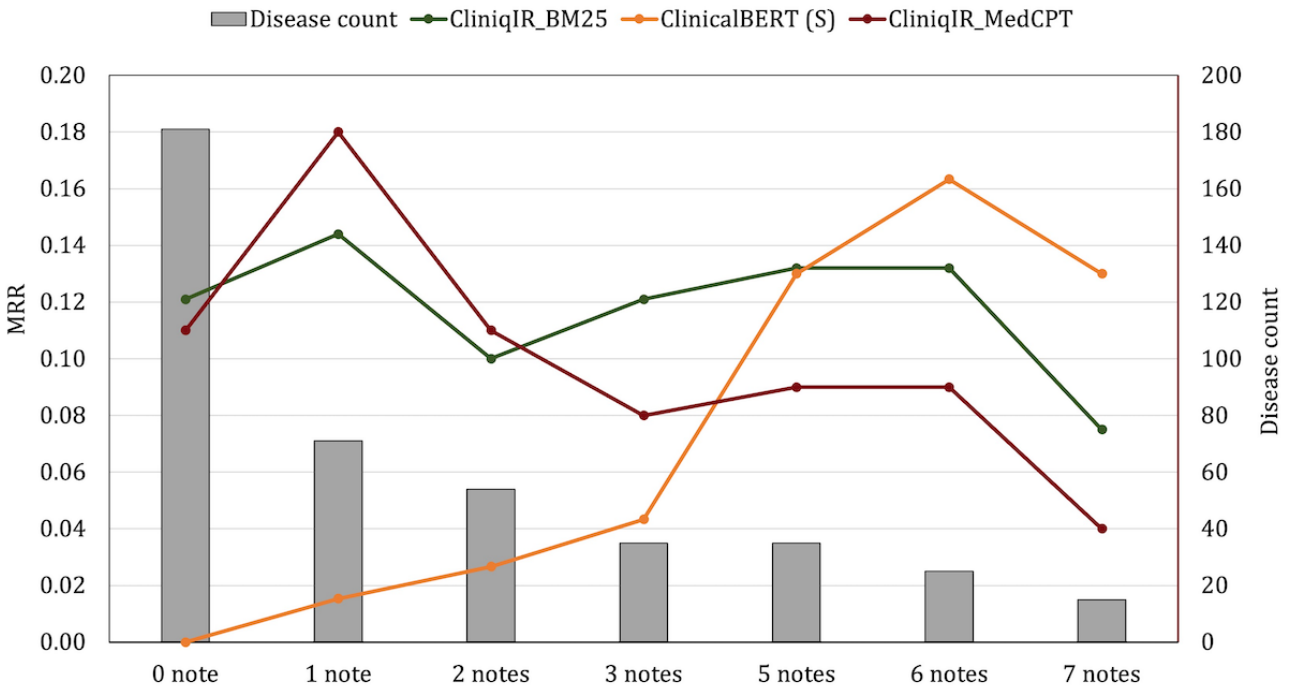
Data set category	Mean reciprocal rank
0 note	—[a]
1≤Notes≤10	0.08
10<Notes≤50	0.33
50<Notes≤100	0.49
100<Notes≤250	0.52
250<Notes≤500	0.57
500<Notes≤750	0.44
750<Notes	0.41
0<Notes	0.37

[a]Not applicable.

CliniqIR Models Outperformed ClinicalBERT for Rare Diagnoses

The objective of this experiment was to determine to what extent the CliniqIR models can be used in place of a supervised model when the training sample size is small. Results in Figure 4 show that CliniqIR-based models performed better than ClinicalBERT for diagnoses with up to 3 training samples. In addition, CliniqIR_BM25 and ClinicalBERT had similar MRR scores for diagnoses with 5 training samples. The average MRR score for the CliniqIR-based models was approximately 0.1 across most categories except for diagnoses with at least 7 training samples. This result indicates that, on average, their correct prediction for a query was ranked 10th on the list. The disease count bars in Figure 4 (in gray) also show that the number of diseases with <5 training samples was more than twice the number of diseases with >5 training samples. Thus, CliniqIR allows for more disease coverage and also generalizes well for cases with low note availability. This result confirms that, while supervised models may perform well with sufficient labeled training data, CliniqIR-based models' performance stands out as remarkable for diagnoses in the low-data regime.

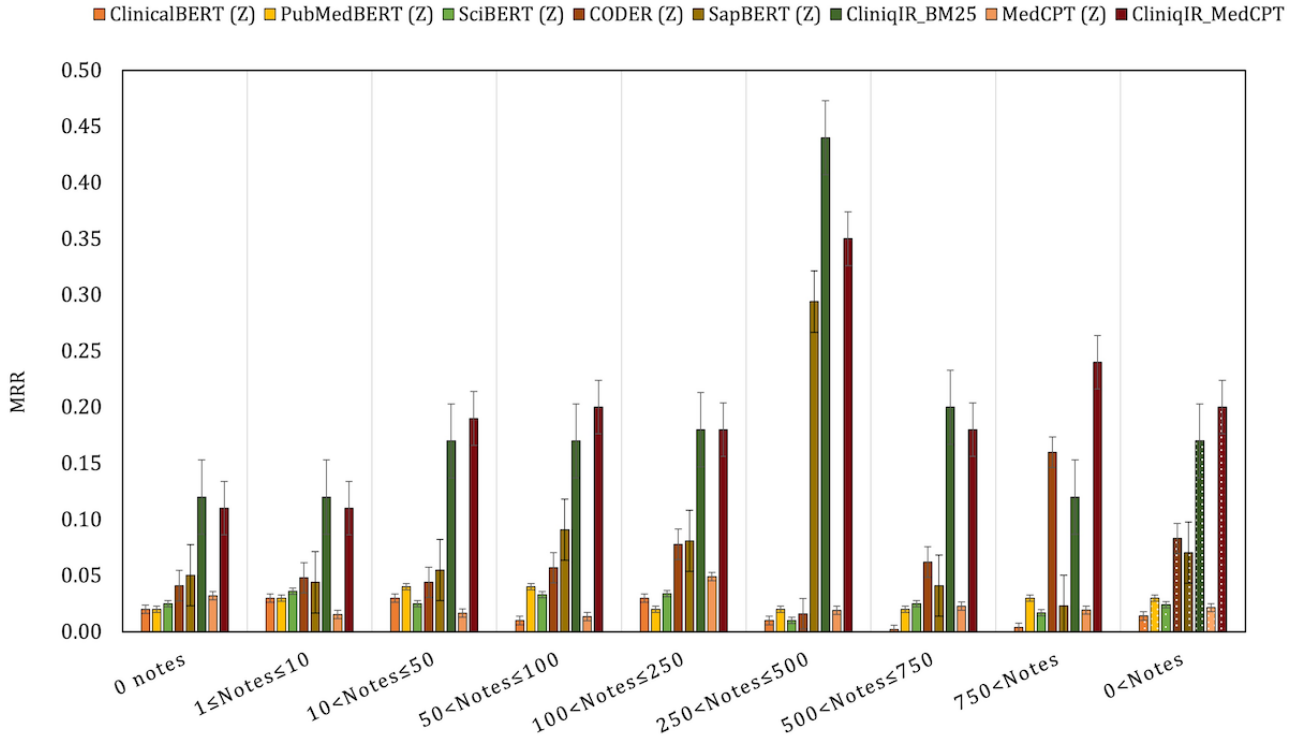
Figure 4. Mean reciprocal rank (MRR) results for CliniqIR-based models and Clinical Bidirectional Encoder Representations from Transformers (ClinicalBERT) when predicting diagnoses with training sample sizes of 0, 1, 2, 3, 5, 6, and 7. Results indicate that the CliniqIR-based models perform best when the training sample size is between 0 and 5. However, ClinicalBERT performs best as training data size increases. "S" denotes that the ClinicalBERT model was used in a supervised setting.



CliniqIR Models Outperformed Zero-Shot Baselines for Rare Diagnoses

To further demonstrate the utility of CliniqIR models for diagnoses with little or no training samples, we compared its performance to that of the pretrained models in a zero-shot setting. As mentioned previously, the MIMIC-III data set comprises >2634 diagnoses, but the supervised fine-tuned models were effective only for a subset of diagnoses with training data; 902 diagnoses had no training samples at all. In zero-shot settings, pretrained models can make predictions without reliance on training data. In Figure 5, we observe that CliniqIR models outperformed the zero-shot pretrained models across most data set categories, especially when diagnoses had a low number of associated training notes. The highest and lowest MRR scores obtained by CliniqIR_BM25 were 0.44 and 0.12, respectively, whereas CliniqIR_MedCPT's highest and lowest scores were 0.35 and 0.11.

Figure 5. Performance evaluation of CliniqIR models and each pretrained zero-shot baseline on the Medical Information Mart for Intensive Care III data set. We categorized the results by the frequency of note representative per diagnosis. "Z" represents models used in a zero-shot setting. The CliniqIR models performed best across data set categories in the low-resource regime. ClinicalBERT: Clinical Bidirectional Encoder Representations from Transformers; CODER: cross-lingual knowledge-infused medical term embeddin; MedCPT: Medical Contrastive Pre-trained Transformers; MRR: mean reciprocal rank; PubMedBERT: PubMed Bidirectional Encoder Representations from Transformers; SapBERT: Self-alignment Pretrained Bidirectional Encoder Representations from Transformers; SciBERT: Scientific Bidirectional Encoder Representations from Transformers.



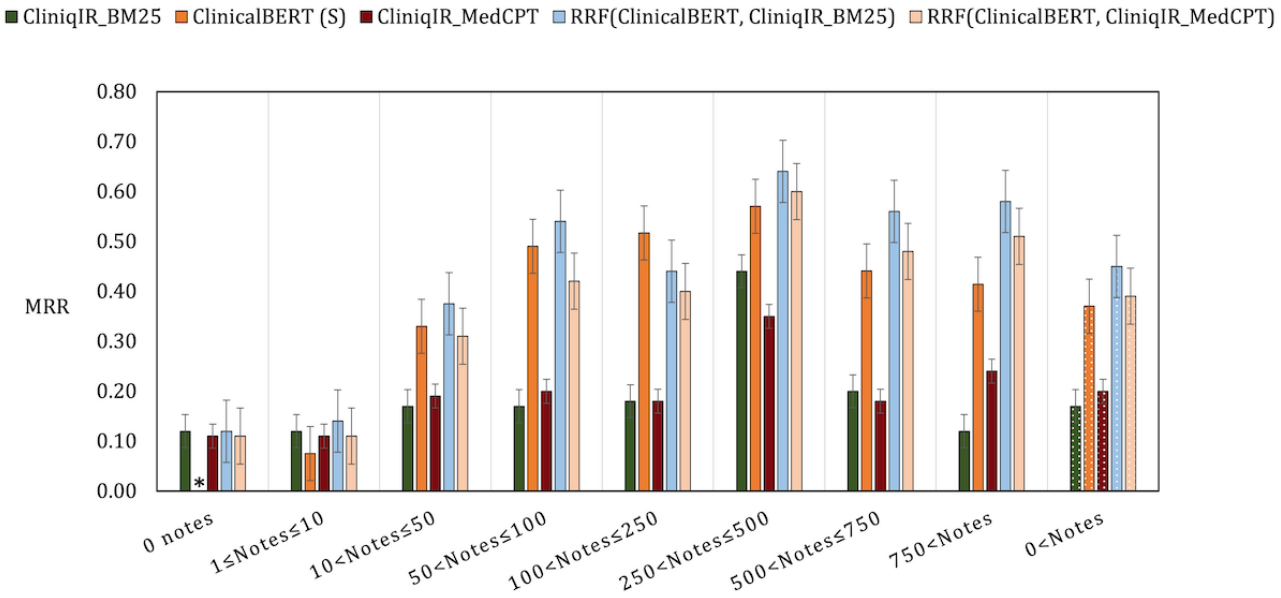
Among the zero-shot baseline, CODER and SapBERT's performances were better across most data set categories. However, in the category in which all diagnoses were considered (diagnoses with >0 notes), CODER outperformed SapBERT, obtaining a maximum and minimum MRR score of 0.16 and 0.02, respectively. These MRR scores indicate that, on average, both CliniqIR models returned the correct diagnosis for a case among their top 5 predictions. In contrast, the best-performing pretrained zero-shot baselines, CODER and SapBERT, returned an accurate diagnosis for a query among their top 15 and 12 predictions, respectively.

Ensemble Models Yielded State-of-the-Art Performance

We have shown that CliniqIR models deliver valuable diagnostic decision support in the setting of limited or unavailable training data. On the other hand, a supervised pretrained model such as ClinicalBERT is an efficient alternative when training data are abundant. To combine the strengths of both models, we used the RRF algorithm as an ensemble strategy. The RRF algorithm combines the ranks of all the ensemble members (a CliniqIR model and a supervised model) to produce a new ranked list of diagnoses for a given patient's clinical note. We hypothesized that creating an ensemble with both models would boost predictive performance across various diagnoses regardless of the availability of associated clinical notes.

To implement the RRF algorithm introduced in the *Model Ensemble: Reciprocal Rank Fusion* section, we used ClinicalBERT and a CliniqIR model to obtain separate ranked lists for each diagnosis and concept across queries. We compared the predictive performance of each individual model to that of the ensemble in terms of MRR for each note availability category. Figure 6 shows the output of the experiments before and after fusing the predicted ranks from both models.

Figure 6. Performance evaluation of the models on the Medical Information Mart for Intensive Care III data set before and after the ensemble. Adopting the reciprocal rank fusion (RRF) algorithm as an ensemble strategy boosted predictive performance across the data set. The Clinical Bidirectional Encoder Representations from Transformer (ClinicalBERT) model cannot directly make predictions for diagnoses with no training samples. Hence, we used "*" to mark such data set categories. The letter "S" denotes that ClinicalBERT was used as a supervised model. MedCPT: Medical Contrastive Pre-trained Transformers; MRR: mean reciprocal rank.



Interestingly, for either of the CliniqIR models used, the ensemble model improved the overall average performance for predicting a wide range of diagnoses (>0 notes) in the MIMIC-III data set. We also found that the RRF ensemble successfully boosted performance across diagnosis categories with both high and low note availability. On average, the RRF ensemble model performed better than either of its constituent models. Notable exceptions include the categories in which the individual mean average precision of both CliniqIR_BM25 and CliniqIR_MedCPT was <0.50 (refer to Multimedia Appendix 1 for details) and in the 100 to 250 training example range, in which the ensemble was slightly worse than the supervised model. In all other conditions, the interaction between both models (the ensemble) led to better performance.

Discussion

Principal Findings

With thousands of known diseases potentially causing a patient's condition, it is often difficult—even for experienced clinicians—to accurately diagnose every disease. Unlike the pretrained models that require user input of possible diagnoses before predictions can be made, CliniqIR represents a potential decision support tool that takes advantage of the wealth of medical literature in PubMed to generate a differential diagnosis. Our study evaluated CliniqIR's ability to formulate differentials and predict uncommon diagnoses with few or no training examples, reflecting conditions easily missed in real-life practice. Results comparing CliniqIR's performance to those of pretrained biomedical transformers in supervised and zero-shot settings highlight CliniqIR's ability to operate successfully as an unsupervised model. Therefore, our model's strength is not limited to rare and infrequent diagnosis prediction, and our model is also a useful tool for generating a first-stage differential diagnosis list. As such, a diagnostic decision support tool such as CliniqIR can enhance physician differential diagnoses and facilitate more efficient diagnoses by providing literature-guided suggestions. Beyond disease prediction, CliniqIR also demonstrates relevance in medical education as a patient-centric literature search tool. Our study demonstrated its ability to accurately cultivate a list of PubMed literature relevant to a patient's clinical narrative. This functionality could greatly improve physician researcher efficiency in performing dedicated literature reviews on behalf of their patients.

In the era of large complex neural models, it is critically important that diagnostic support tools remain simple and interpretable. In health care, where decision-making is critical and patient outcomes are at stake, clinicians' ability to understand and trust the inner workings of a diagnostic tool is paramount. In response, CliniqIR is built on retrieval systems that use simple and transparent weighting schemes to retrieve and rank important terms in a collection of documents. This transparency fosters trust in the tool's accuracy and facilitates collaboration between the tool and the health care professionals, leading to ongoing model refinement as well as enhanced clinical decision-making.

Limitations

The medical field is witnessing a growing trend in applications built on generative large language models [52]. While our work used a simpler approach, it remains valuable in scenarios with limited access to significant computing resources. In addition, it serves as a proof of concept for a retrieval-augmented medical model, potentially leading to enhanced explainability and accuracy for large language models in the health care domain.

Our study has 3 potential limitations. First, CliniqIR's knowledge source is limited to abstracts in PubMed, which has well-known publication biases toward certain diagnoses [53,54].

Therefore, the use of a single knowledge resource limits CliniqIR's generalizability to diseases and conditions not represented in the PubMed corpus. For instance, conditions such as COVID-19 and Alzheimer disease or rare diseases such as sarcoidosis and cholangitis are covered in thousands of published literature entries, whereas other conditions such as "cellulitis and abscess of the leg" or "closed fracture of the sternum" may receive less attention. Future studies will involve a review of seemingly unrepresented diagnostic codes by linking them back to their parent diagnostic codes to ensure appropriate mapping between diagnosis codes and PubMed.

Second, our main experimental results were restricted to diagnoses with at least 100 PubMed abstract representatives. We identified a significant number of *ICD-9-CM* codes in MIMIC-III with no associated medical literature among the 33 million PubMed abstracts (an overview of MIMIC-III diagnosis distribution classes can be found in Multimedia Appendix 1). We also found that CliniqIR's predictive performance improved with increasing PubMed coverage of the diagnosis, guiding our decision to establish the 100-abstract inclusion criterion for diagnoses (Multimedia Appendix 1). Future work will combine information from biomedical journals, medical textbooks, and Wikipedia for wider disease coverage.

Third, our MIMIC-III experiments limited the input to patient discharge summaries containing a succinct synopsis of a patient's hospital stay, including symptoms, diagnostic evaluation, clinical progression, and treatment information. In real-world clinical situations, such complete retrospective information would not be available during the initial diagnostic process. Therefore, the results presented in this paper represent a first feasibility study of CliniqIR and highlight some of the difficulties involved in developing diagnostic support tools.

Conclusions

In this study, we presented CliniqIR, an unsupervised retrieval-based model that leverages unstructured knowledge resources to aid in the diagnostic process. We showed that the CliniqIR models outperformed a supervised fine-tuned pretrained clinical transformer model in predicting diagnoses with <5 training samples. We also demonstrated that CliniqIR outperformed pretrained clinical transformers in making predictions for rare and complex conditions in a zero-shot setting. While many existing research studies on diagnostic prediction have focused on one disease at a time or only on highly prevalent conditions, we combined the strengths of CliniqIR and supervised learning to build a single ensemble model that aids in diagnosing a broad spectrum of conditions regardless of training data availability. Overall, our study reveals the potential of IR-based models in aiding diagnostic decision-making in an efficient, transparent, and educational manner. This work will direct future studies to facilitate successful application of machine learning and IR to building robust and accurate clinical diagnostic decision support tools.

Acknowledgments

LM's research is supported in part by a grant (T32DA013911) from the National Institutes of Health. We acknowledge support from the Open Access Publication Fund of the University of Tübingen.

Data Availability

The datasets analyzed during this study are available. The Medical Information Mart for Intensive Care III data set can be accessed with permission via the work by Johnson et al [39]. The DC3 data set can be accessed via the work by Eickhoff et al [30]. The code to implement this work is also available [55].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details on model implementation, configuration, and performance comparisons across various data sets and configurations.
[DOCX File , 5518 KB - medinform_v12i1e50209_app1.docx]

References

1. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain: an evolving diagnostic decision-support system. JAMA 1987 Jul 03;258(1):67-74. [doi: 10.1001/jama.1987.03400010071030]
2. Khoong EC, Nouri SS, Tuot DS, Nundy S, Fontil V, Sarkar U. Comparison of diagnostic recommendations from individual physicians versus the collective intelligence of multiple physicians in ambulatory cases referred for specialist consultation. Med Decis Making 2022 Apr;42(3):293-302 [FREE Full text] [doi: 10.1177/0272989X211031209] [Medline: 34378444]
3. Barnett ML, Boddupalli D, Nundy S, Bates DW. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. JAMA Netw Open 2019 Mar 01;2(3):e190096 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.0096] [Medline: 30821822]
4. 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 06;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]

5. Ramnarayan P, Tomlinson A, Rao A, Coren M, Winrow A, Britto J. ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation. Arch Dis Child 2003 May;88(5):408-413 [FREE Full text] [doi: 10.1136/adc.88.5.408] [Medline: 12716712]
6. Müller L, Gangadharaiah R, Klein SC, Perry J, Bernstein G, Nurkse D, et al. An open access medical knowledge base for community driven diagnostic decision support system development. BMC Med Inform Decis Mak 2019 Apr 27;19(1):93 [FREE Full text] [doi: 10.1186/s12911-019-0804-1] [Medline: 31029130]
7. Prakash A, Zhao S, Hasan S, Datla V, Lee K, Qadir A, et al. Condensed memory networks for clinical diagnostic inferencing. Proc AAAI Conf Artif Intell 2017 Feb 12;31(1). [doi: 10.1609/aaai.v31i1.10964]
8. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Sci Rep 2021 Feb 05;11(1):3254 [FREE Full text] [doi: 10.1038/s41598-020-74399-w] [Medline: 33547343]
9. Liu J, Zhang Z, Razavian N. Deep EHR: chronic disease prediction using medical notes. arXiv Preprint posted online August 15, 2018 [FREE Full text]
10. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One 2018 Feb 15;13(2):e0192360 [FREE Full text] [doi: 10.1371/journal.pone.0192360] [Medline: 29447188]
11. Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: a review. Egypt Inform J 2018 Nov;19(3):179-189. [doi: 10.1016/j.eij.2018.03.002]
12. Singh Kohli P, Arora S. Application of machine learning in disease prediction. In: Proceedings of the 4th International Conference on Computing Communication and Automation. 2018 Presented at: ICCCA 2018; December 14-15, 2018; Greater Noida, India. [doi: 10.1109/ccaa.2018.8777449]
13. Abdullahi T, Nitschke G, Sweijd N. Predicting diarrhoea outbreaks with climate change. PLoS One 2022 Apr 19;17(4):e0262008 [FREE Full text] [doi: 10.1371/journal.pone.0262008] [Medline: 35439258]
14. Alon G, Chen E, Savova G, Eickhoff C. Diagnosis prevalence vs. efficacy in machine-learning based diagnostic decision support. arXiv Preprint posted online June 24, 2020 [FREE Full text]
15. Rios A, Kavuluru R. Few-shot and zero-shot multi-label learning for structured label spaces. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP 2018; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-1352]
16. Zhao Y, Wong ZS, Tsui KL. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. J Healthc Eng 2018 May 22;2018:6275435 [FREE Full text] [doi: 10.1155/2018/6275435] [Medline: 29951182]
17. Romera-Paredes B, Torr PH. An embarrassingly simple approach to zero-shot learning. In: Feris R, Lampert C, Parikh D, editors. Visual Attributes. Cham, Switzerland: Springer; Mar 22, 2017.
18. Veeranna SP, Nam J, Mencía EL, Furnkranz J. Using semantic similarity for multi-label zero-shot classification of text documents. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. 2016 Presented at: ESANN 2016; April 27-29, 2016; Bruges, Belgium.
19. Brown TB, 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]
20. Sanh V, Webson A, Raffel C, Bach SH, Sutawika L, Alyafeai Z, et al. Multitask prompted training enables zero-shot task generalization. arXiv Preprint posted online October 15, 2021 [FREE Full text]
21. Li Y, Chao X. Semi-supervised few-shot learning approach for plant diseases recognition. Plant Methods 2021 Jun 27;17(1):68 [FREE Full text] [doi: 10.1186/s13007-021-00770-1] [Medline: 34176505]
22. Zhao Y, Lai H, Yin J, Zhang Y, Yang S, Jia Z, et al. Zero-shot medical image retrieval for emerging infectious diseases based on meta-transfer learning - worldwide, 2020. China CDC Wkly 2020 Dec 25;2(52):1004-1008 [FREE Full text] [doi: 10.46234/ccdcw2020.268] [Medline: 34594825]
23. Soldaini L, Cohan A, Yates A, Goharian N, Frieder O. Retrieving medical literature for clinical decision support. In: Proceedings of the 37th European Conference on IR Research. 2015 Presented at: ECIR 2015; March 29-April 2, 2015; Vienna, Austria. [doi: 10.1007/978-3-319-16354-3_59]
24. Hasan SA, Ling Y, Liu J, Farri O. Using neural embeddings for diagnostic inferencing in clinical question answering. In: Proceedings of the Twenty-Fourth Text REtrieval Conference. 2015 Presented at: TREC 2015; November 17-20, 2015; Gaithersburg, MD.
25. Hasan SA, Zhu X, Dong Y, Liu J, Farri O. A hybrid approach to clinical question answering. In: Proceedings of the Twenty-Third Text REtrieval Conference 2014. 2014 Presented at: TREC 2014; November 19-21, 2014; Gaithersburg, MD.
26. Text REtrieval Conference (TREC) home page. Text REtrieval Conference (TREC). URL: https://trec.nist.gov/ [accessed 2024-06-03]
27. Naik A, Parasa S, Feldman S, Wang LL, Hope T. Literature-augmented clinical outcome prediction. arXiv Preprint posted online November 16, 2021 [FREE Full text] [doi: 10.18653/v1/2022.findings-naacl.33]
28. 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]

29. White J. PubMed 2.0. Med Ref Serv Q 2020 Oct 21;39(4):382-387. [doi: 10.1080/02763869.2020.1826228] [Medline: 33085945]
30. Eickhoff C, Gmehlin F, Patel AV, Boullier J, Fraser H. DC3 -- a diagnostic case challenge collection for clinical decision support. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. 2019 Presented at: ICTIR '19; October 2-5, 2019; Santa Clara, CA. [doi: 10.1145/3341981.3344239]
31. Soldaini L, Goharian N. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In: Proceedings of the 2nd SIGIR Workshop on Medical Information Retrieval (MedIR). 2016 Presented at: SIGIR 2016; July 21, 2016; Pisa, Italy.
32. Zhao Z, Jin Q, Chen F, Peng T, Yu S. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. Sci Data 2023 Dec 18;10(1):909 [FREE Full text] [doi: 10.1038/s41597-023-02814-8] [Medline: 38110415]
33. Vardell E, Moore M. Isabel, a clinical decision support system. Med Ref Serv Q 2011 Apr 25;30(2):158-166. [doi: 10.1080/02763869.2011.562800] [Medline: 21534115]
34. Welcome to Apache Lucene. Apache Lucene. URL: https://lucene.apache.org/ [accessed 2024-06-03]
35. Robertson S, Walker S, Hancock-Beaulieu MM, Gatford M, Payne A. Okapi at TREC-4. In: Proceedings of the Fourth Text REtrieval Conference. 1995 Presented at: TREC-4 1995; November 1-3, 1995; Gaithersburg, MD.
36. Jin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Wilbur WJ, et al. MedCPT: contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. Bioinformatics 2023 Nov 01;39(11):2023 [FREE Full text] [doi: 10.1093/bioinformatics/btad651] [Medline: 37930897]
37. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP 2019; June 7, 2019; Minneapolis, MN. [doi: 10.18653/v1/w19-1909]
38. 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]
39. Johnson AE, Pollard TJ, Shen LW, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
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 Healthcare 2021 Oct 15;3(1):1-23. [doi: 10.1145/3458754]
41. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. arXiv Preprint posted online March 26, 2019 [FREE Full text] [doi: 10.18653/v1/d19-1371]
42. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. arXiv Preprint posted online October 22, 2020 [FREE Full text] [doi: 10.18653/v1/2021.naacl-main.334]
43. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: knowledge-infused cross-lingual medical term embedding for term normalization. J Biomed Inform 2022 Feb;126:103983 [FREE Full text] [doi: 10.1016/j.jbi.2021.103983] [Medline: 34990838]
44. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv Preprint posted online August 27, 2019. [doi: 10.18653/v1/d19-1410]
45. Cormack GV, Clarke CL, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009 Presented at: SIGIR '09; July 19-23, 2009; Boston, MA. [doi: 10.1145/1571941.1572114]
46. Kurland O, Culpepper J. Fusion in information retrieval: SIGIR 2018 half-day tutorial. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018 Presented at: SIGIR '18; July 8-12, 2018; Ann Arbor, MI. [doi: 10.1145/3209978.3210186]
47. Mugisha C, Paik I. Comparison of neural language modeling pipelines for outcome prediction from unstructured medical text notes. IEEE Access 2022;10:16489-16498. [doi: 10.1109/access.2022.3148279]
48. Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J Am Med Inform Assoc 2019 Dec 01;26(12):1632-1636 [FREE Full text] [doi: 10.1093/jamia/ocz164] [Medline: 31550356]
49. Shen Z, Schutte D, Yi Y, Bompelli A, Yu F, Wang Y, et al. Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. BMC Med Inform Decis Mak 2022 Jul 07;22(Suppl 1):88 [FREE Full text] [doi: 10.1186/s12911-022-01819-4] [Medline: 35799294]
50. Craswell N. Mean reciprocal rank. In: Liu L, Özsu MT, editors. Encyclopedia of Database Systems. Boston, MA: Springer; 2009.
51. Basic HHS Policy for Protection of Human Research Subjects. US Department of Health and Human Services, Office for Human Research Protections. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/regulatory-text/index.html#46.101
52. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. JMIR Med Educ 2024 Feb 13;10:e51391 [FREE Full text] [doi: 10.2196/51391] [Medline: 38349725]

53. McGauran N, Wieseler B, Kreis J, Schüler YB, Kölsch H, Kaiser T. Reporting bias in medical research - a narrative review. Trials 2010 Apr 13;11(1):37 [FREE Full text] [doi: 10.1186/1745-6215-11-37] [Medline: 20388211]
54. Montori VM, Smieja M, Guyatt GH. Publication bias: a brief review for clinicians. Mayo Clin Proc 2000 Dec;75(12):1284-1288. [doi: 10.4065/75.12.1284] [Medline: 11126838]
55. CliniqIR: retrieval-based diagnostic decision support. GitHub. URL: https://github.com/rsinghlab/CliniqIR [accessed 2024-05-31]

Abbreviations

BERT: Bidirectional Encoder Representations From Transformers
ClinicalBERT: Clinical Bidirectional Encoder Representations from Transformers
CODER: cross-lingual knowledge-infused medical term embedding
DDSS: diagnostic decision support system
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
IR: information retrieval
MAP: mean average precision
MedCPT: Medical Contrastive Pre-trained Transformers
MIMIC-III: Medical Information Mart for Intensive Care III
MRR: mean reciprocal rank
PubMedBERT: PubMed Bidirectional Encoder Representations from Transformers
RR: reciprocal rank
RRF: reciprocal rank fusion
SapBERT: Self-alignment Pretrained Bidirectional Encoder Representations from Transformers
SciBERT: Scientific Bidirectional Encoder Representations from Transformers
TF: term frequency
TF-IDF: term frequency–inverse document frequency
TREC: text retrieval conference
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 25.06.23; peer-reviewed by J Zheng, Q Jin; comments to author 06.02.24; revised version received 10.03.24; accepted 17.04.24; published 19.06.24.

Please cite as:
Abdullahi T, Mercurio L, Singh R, Eickhoff C
Retrieval-Based Diagnostic Decision Support: Mixed Methods Study
JMIR Med Inform 2024;12:e50209
URL: https://medinform.jmir.org/2024/1/e50209
doi:10.2196/50209
PMID:38896468

An Electronic Health Record–Integrated Application for Standardizing Care and Monitoring Patients With Autosomal Dominant Polycystic Kidney Disease Enrolled in a Tolvaptan Clinic: Design and Implementation Study

Maroun Chedid[1], MD; Fouad T Chebib[2], MD; Erin Dahlen[3], BSN; Theodore Mueller[3], BSN; Theresa Schnell[3], BSN; Melissa Gay[3], BSN; Musab Hommos[4], MBBS; Sundararaman Swaminathan[4], MBBS; Arvind Garg[5], MBBS; Michael Mao[2], MD; Brigid Amberg[3], BSN; Kirk Balderes[6], BS; Karen F Johnson[6], MA; Alyssa Bishop[6], MBA; Jackqueline Kay Vaughn[6], MA; Marie Hogan[3], MD, PhD; Vicente Torres[3], MD, PhD; Rajeev Chaudhry[7], MBBS, MPH; Ziad Zoghby[3], MBA, MD

1

2

3

4

5

6

7

Corresponding Author:
Ziad Zoghby, MBA, MD

Abstract

Background: Tolvaptan is the only US Food and Drug Administration–approved drug to slow the progression of autosomal dominant polycystic kidney disease (ADPKD), but it requires strict clinical monitoring due to potential serious adverse events.

Objective: We aimed to share our experience in developing and implementing an electronic health record (EHR)–based application to monitor patients with ADPKD who were initiated on tolvaptan.

Methods: The application was developed in collaboration with clinical informatics professionals based on our clinical protocol with frequent laboratory test monitoring to detect early drug-related toxicity. The application streamlined the clinical workflow and enabled our nursing team to take appropriate actions in real time to prevent drug-related serious adverse events. We retrospectively analyzed the characteristics of the enrolled patients.

Results: As of September 2022, a total of 214 patients were enrolled in the tolvaptan program across all Mayo Clinic sites. Of these, 126 were enrolled in the Tolvaptan Monitoring Registry application and 88 in the Past Tolvaptan Patients application. The mean age at enrollment was 43.1 (SD 9.9) years. A total of 20 (9.3%) patients developed liver toxicity, but only 5 (2.3%) had to discontinue the drug. The 2 EHR-based applications allowed consolidation of all necessary patient information and real-time data management at the individual or population level. This approach facilitated efficient staff workflow, monitoring of drug-related adverse events, and timely prescription renewal.

Conclusions: Our study highlights the feasibility of integrating digital applications into the EHR workflow to facilitate efficient and safe care delivery for patients enrolled in a tolvaptan program. This workflow needs further validation but could be extended to other health care systems managing chronic diseases requiring drug monitoring.

(JMIR Med Inform 2024;12:e50164) doi:10.2196/50164

KEYWORDS

ADPKD; autosomal dominant polycystic kidney disease; polycystic kidney disease; tolvaptan; EHR; electronic health record; digital health solutions; monitoring; kidney disease; drug-related toxicity; digital application; management; chronic disease

Introduction

Autosomal dominant polycystic kidney disease (ADPKD) is the leading genetic cause and the fourth overall cause of end-stage kidney failure (ESKF) [1]. Patients with polycystic kidney disease 1 (PKD1) mutations develop ESKF 20 years earlier than those with PKD2 mutations [2,3]. The Mayo imaging classification (MIC) is a validated tool that identifies patients at risk for rapid progression to kidney failure, and disease-modifying therapy is recommended for patients with class 1C, 1D, or 1E, who have higher total kidney volume (TKV) growth rates [4]. In 2018, tolvaptan (brand name Jynarque; Otsuka America Pharmaceutical) was approved by the US Food and Drug Administration (FDA) as the first drug to slow kidney function decline in patients with rapidly progressive ADPKD. Tolvaptan reduces kidney volume growth and estimated glomerular filtration rate (eGFR) decline, delaying the need for kidney replacement therapy [5,6]. Tolvaptan acts by blocking the vasopressin V2 receptors in the distal nephron and collecting duct, inhibiting urinary concentration and sodium reabsorption and reversing the tubuloglomerular feedback inhibition induced by vasopressin, thus acutely and reversibly decreasing eGFR and possibly glomerular hyperfiltration [5]. However, tolvaptan is associated with several side effects, including polyuria, urinary frequency, thirst, and nocturia, which require patient education on adequate hydration. Tolvaptan can also cause significant hepatotoxicity in 5% of patients; thus, periodic liver function tests are mandated by the FDA through the risk evaluation and mitigation strategy (REMS) safety program. Due to the side effects profile and the necessary frequent laboratory test monitoring, the cost associated with staff time to manage the program beyond face-to-face care can limit the ability of health care teams to safely provide this disease-modifying therapy [7,8].

Tools that are directly integrated in the electronic health record (EHR) workflow can increase efficiency, reduce cost, and improve drug monitoring and quality of care [9-12]. For example, a cluster randomized clinical trial in primary care provided access, within the EHR, to a prescription drug monitoring program (PDMP) before the prescription of opioids. The integration increased PDMP-querying rates, suggesting that direct access reduced hassle costs and could improve adherence to guideline-concordant care practices [13]. Another study reported that the design and implementation of an electronic registry with a complementary workflow established an active tracking system that improved monitoring of patients on anticoagulation therapy [14]. However, no prior EHR-integrated workflow has been developed and validated to safely and successfully monitor patients with ADPKD treated with tolvaptan.

This paper describes the design, development, and implementation of an intelligent automated application within the EHR to efficiently manage and monitor ADPKD patients enrolled in the Mayo Clinic tolvaptan program. The goal of this paper is to illustrate how digital applications integrated into the EHR workflow can facilitate efficient and safe care for patients enrolled in a drug monitoring program and how this workflow can be extended to similar programs in chronic disease management and lay the groundwork for quality improvement efforts.

Methods

Ethical Considerations

This work was reviewed by the Mayo Clinic Institutional Review Board (21-005428). The study was exempt from clinical research overight because it was considered to be a quality improvement project. Informed consent was waived and data were deidentified.

Practice Setting

The Mayo Clinic is an integrated, multispecialty, multistate, large academic health system with locations in Minnesota, Florida, and Arizona; there are also other Mayo Clinic health system hospitals across Minnesota and Wisconsin. Since 2018, the Mayo Clinic uses a single, integrated EHR (Epic Systems) across all campuses. The Minnesota practice where the tolvaptan EHR application was initially launched includes 5 experienced nephrologists and 4 nurses directly involved in the ADPKD practice and various other specialists (ie, geneticists, hepatologists, liver surgeons, neurologists, neurosurgeons, pain specialists, interventional radiologists, transplant experts, and research coordinators) who care for these patients as needed. The tolvaptan EHR application was eventually adopted enterprise-wide in 2022, although the workflow may differ slightly by site based on the specificity and resources available in each practice. The 3 main Mayo Clinic campuses in Minnesota, Florida, and Arizona are designated as centers of excellence for ADPKD care by the Polycystic Kidney Disease Foundation.

Clinical Protocol—Indications and Monitoring of Tolvaptan Treatment

Tolvaptan is prescribed in patients aged 18 to 55 years with eGFR ≥25 mL/min/1.73 m^2 and at risk of rapid progression, defined by having an age-indexed height-adjusted TKV within MIC class 1C, 1D, and 1E [5,6,15]. Contraindications to initiate tolvaptan include history of liver injury, uncorrected hypernatremia, hypovolemia, inability to sense thirst, urinary tract obstruction, and concomitant use of strong CYP3A (cytochrome P450, family 3, subfamily A) enzyme inhibitors [16]. Tolvaptan initiation requires a multidisciplinary approach led by the treating nephrologist and a well-trained nursing team. In our program, following a shared decision discussion, eligible patients who agree to start tolvaptan are referred to a specialized nephrology nurse for a detailed educational session. The nurse visit includes a blood pressure check, assessment of alcohol consumption, dietary review, and in-depth education about the side effects of tolvaptan and the need for routine laboratory test monitoring. Patients are instructed to have a drug holiday in certain situations, such when they are unable to maintain adequate fluid intake, are hospitalized, are about to undergo an elective procedure, or are traveling. After confirming their willingness to take the medication, the nurse enrolls the patient in the mandatory REMS program, a drug safety program developed by the FDA for certain medications with serious safety concerns. As part of the tolvaptan REMS program, the

following laboratory tests are performed before the morning dose of tolvaptan: aspartate transaminase (AST), alanine transaminase (ALT), total bilirubin, serum sodium (advised but optional), and creatinine (advised but optional). Results are collected 2 and 4 weeks after tolvaptan initiation, then monthly for 18 months, and every 3 months thereafter [16]. Staff must log in to complete a REMS attestation every 3 months for each patient. Liver enzyme elevation and changes in serum sodium or creatinine are reviewed after each test in a timely fashion by the nursing team and the prescribing nephrologist. This process is designed to detect any laboratory test abnormality or the development of drug complications that could otherwise go unnoticed. For example, one threshold for suspending the medication is elevation in AST or ALT twice above their baseline level, which might not be flagged in the test report. However, this process can be very cumbersome and time consuming for the clinical team. An intelligent, automated, and streamlined real-time EHR-based process of tracking and monitoring is essential for efficient and safe care delivery, especially in specialized centers with a large patient population.

Architecture and Application Development by the Cohort Knowledge Intelligence Solutions Team

At the Mayo Clinic, the Cohort Knowledge Intelligence Solutions (CKIS) team is behind the development of many innovative patient cohort management solutions using the Epic Healthy Planet module. The CKIS team uses a collaborative, agile approach that incorporates feedback from clinical stakeholders and informatics to create care management solutions based on agreed-upon protocols of care that improve and automate processes for clinical staff, all managed within the EHR. All projects are reviewed through a standard intake process that factors in the scope of the project, enterprise impact, patient safety, quality of care, and revenue impact, among other criteria. Once approved and assigned, a business analyst and a builder will work with a group of stakeholders anywhere between several weeks to several months, depending on the scope of the project, to complete a solution build.

After the scope of a project is defined, a registry is used to gather a patient cohort along with a subset of metrics required to support the practice needs. The registry is an internal tool housed within Epic's software and uses a rule-based framework consisting of 2 main components: an inclusion rule and metrics. The inclusion rule is used to define the population and uses a combination of charted data, such as the patient's diagnosis, medication, and surgeries, or general demographics (eg, age and gender). The metrics (ie, rules) define what data will be captured for the population. Once a patient meets the defined inclusion criteria, the underlying metrics are processed, and data is captured. Metrics are designed to support the monitoring workflow in addition to future quality analysis and outcomes. They typically capture dates, laboratory test values, appointment information, patient demographics, and more. Finally, a report is built allowing users to visualize and interact with the registry data. Within the reports, specific patient metrics are displayed pertinent to the practice and may include laboratory test results, appointment information, or customized algorithms to create alerts for care team members to help them prioritize their work. The last phase of the build process includes testing and ensuring that the initial agreed-upon requirements have been met. Several months after the build is complete, the CKIS team meets with the customers to complete a value assessment that measures the impact of the solution provided.

Process of Tolvaptan Application Development

The nephrology ADPKD practice assembled a team of stakeholders to streamline the enrollment and monitoring of patients in the tolvaptan program. After an initial discussion in early 2020, the clinical team determined the content of the application. The stakeholders met on average every 2 weeks over a 3-month period to develop, in an iterative fashion, the initial application and test the efficiency of the system over the subsequent 3 to 4 months. The team determined that 2 applications were required to serve the clinical need. The first and main application, titled Tolvaptan Monitoring Registry, manages all patients actively treated with tolvaptan by consolidating all relevant information in one screen. Patients who discontinue tolvaptan are removed from the first application and automatically added to the second application, titled Past Tolvaptan Patients. The second application allows the care team to maintain a log of all past participants and record the reason for drug discontinuation, such as adverse effects or requiring renal replacement therapy.

Results

In September 2020, 2 EHR-based applications for monitoring patients taking tolvaptan were activated for clinical use. The tolvaptan clinic was established 2 years earlier when the FDA approved tolvaptan for the treatment of ADPKD. All patients enrolled in the program prior to September 2020 (n=32) were retrospectively added to the tolvaptan monitoring application.

Clinical Workflow Using the Tolvaptan Application

The tolvaptan application workflow involves the submission of an electronic prescription order by a nephrology nurse (Figure 1) and completion of an electronic activation form to enroll patients in the EHR-based tolvaptan monitoring application (Figure 2). This form includes the patient's Mayo Clinic site, primary nephrology clinician, and date of treatment initiation. The automated addition of patients into the registry reduces the risk of missing any patient prescribed tolvaptan, ensuring that all treated patients are closely monitored for any adverse events that might occur while on therapy. Quarterly meetings take place between the nursing team and the nephrologists to review workflow issues and assess any new complications that may arise.

Figure 1. Tolvaptan order report. REMS: risk evaluation and mitigation strategy.

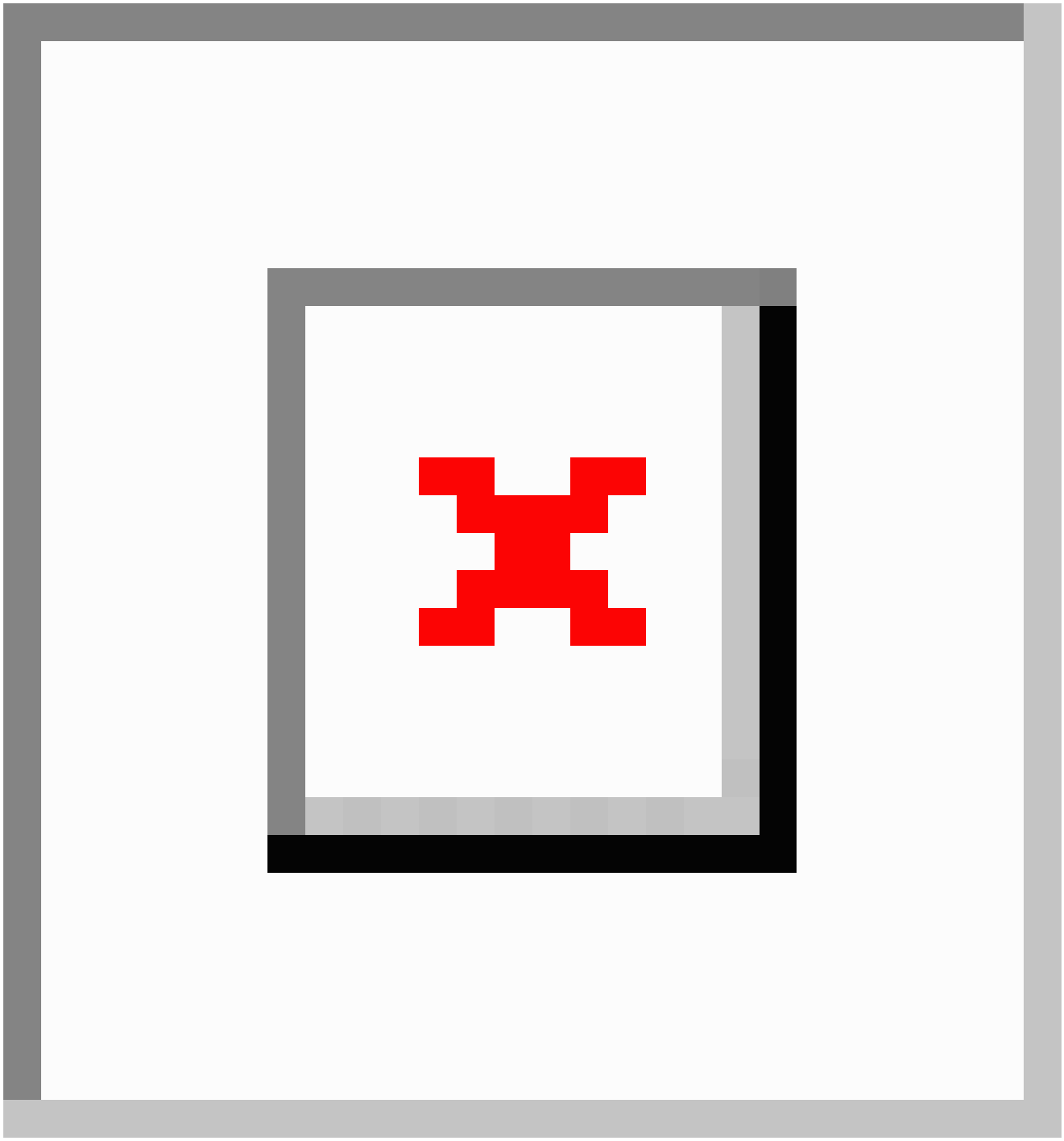
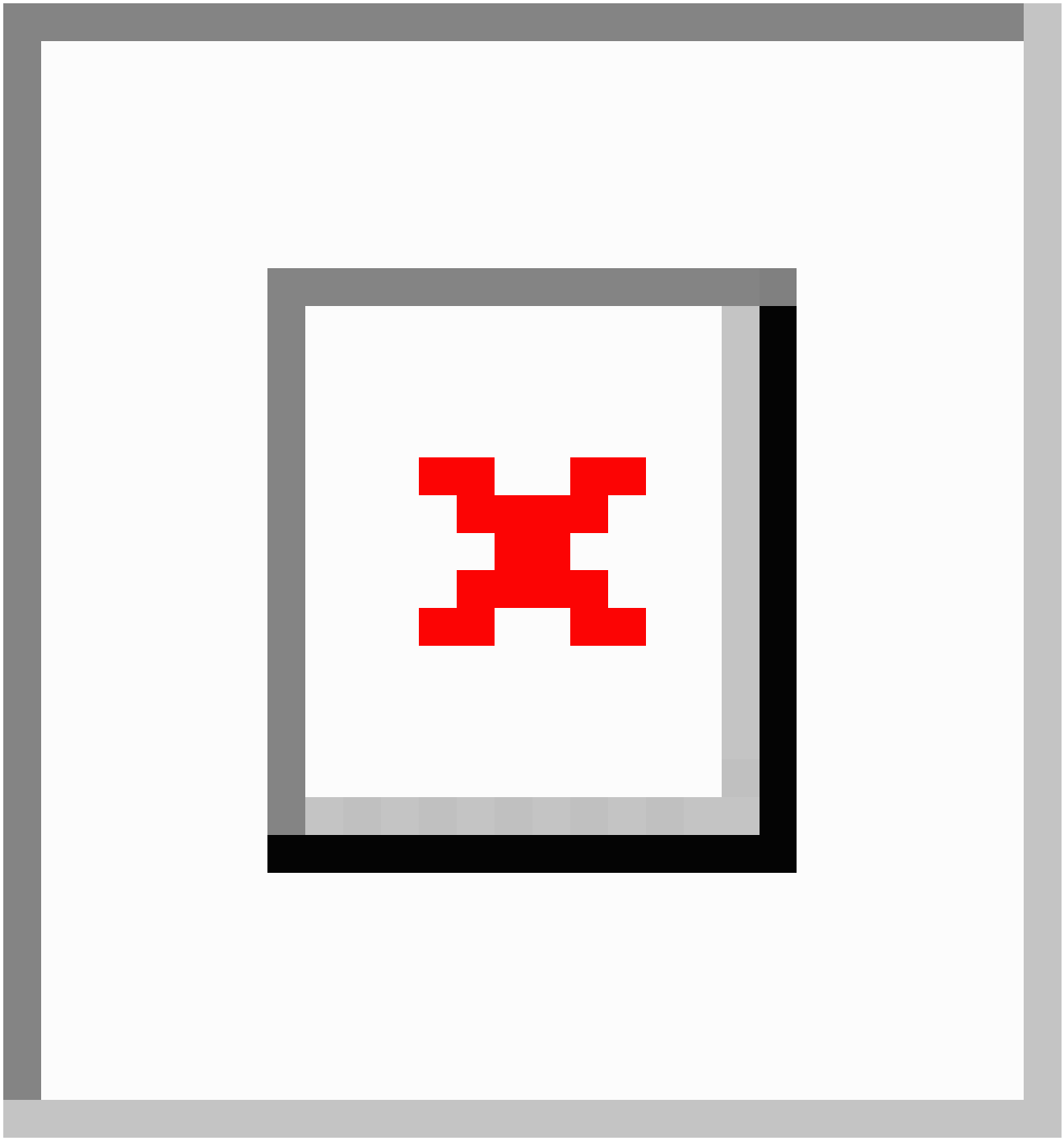


Figure 2. Smart form. MCHS: Mayo Clinic Health System; MN: Minnesota; NE: northeast; NW: northwest; REMS: risk evaluation and mitigation strategy; SW: southwest; WI: Wisconsin.



The main application lists all patients actively taking tolvaptan and includes several columns with relevant information, such as the patient's name, medical record number, date of initiation of tolvaptan, date of last liver function test, abnormal or urgent laboratory test flags, a "needs review" flag (Multimedia Appendix 1 provides details and flag criteria), the recommended laboratory test frequency based on the first dose date (ie, monthly or quarterly), the laboratory test's due date and the date of the next scheduled laboratory test, the last and next (when applicable) nursing outreach dates to the patient, the name of the treating nephrologist, and the last clinic visit date (Figure 3). The application allows filtering based on these variables, such as visualizing only patients who have abnormal laboratory tests or need review based on new laboratory tests since the last outreach date. The application also provides more detailed information for a specific patient based on several reports embedded at the bottom of the screen. These include Tolvaptan Monitoring Summary, Patient Visits, Nephrology Notes/Orders, and Patient Message Review. In our clinical workflow, every week, 1 of 4 dedicated nurses (on a rotation basis) reviews all flagged patients and takes appropriate action based on our clinical protocol. The EHR-based tolvaptan monitoring application provides several reports that allow for a more detailed review of a specific patient without having to open their chart. The Tolvaptan Monitoring Summary displays all monitored laboratory test results, such as AST, ALT,

bilirubin, serum creatinine, eGFR, serum sodium, and urine osmolarity (Figure 4). The Tolvaptan Monitoring Metrics window in the same section allows for quick access to recorded baseline laboratory test measurements and any abnormalities recorded. For example, the report displays a question and response: “Any Abnormal Liver Labs?” (answers are yes or no) (Figure 5). Additionally, all attempted or completed outreach interactions are listed with the name of the nurse conducting the activity and most recent nephrology note (Figure 6).

The Patient Visits report shows future scheduled appointments and surgeries, as well as a record of the patient’s last 10 outpatient visits. This report also includes the patient’s care team, demographics, and emergency contacts. The Nephrology Notes/Orders report displays pertinent medical history, current medication, immunizations, renal replacement therapy status, allergies, procedures, and the latest nephrology notes and specific ADPKD management–related comments by the nephrologist. Lastly, the Patient Message Review report includes all the patient’s communications with personnel, nurses, and clinicians, as well as patient online services.

Figure 3. Tolvaptan monitoring snapshot. Abn: abnormal; Dt: date; REMS: risk evaluation and mitigation strategy.

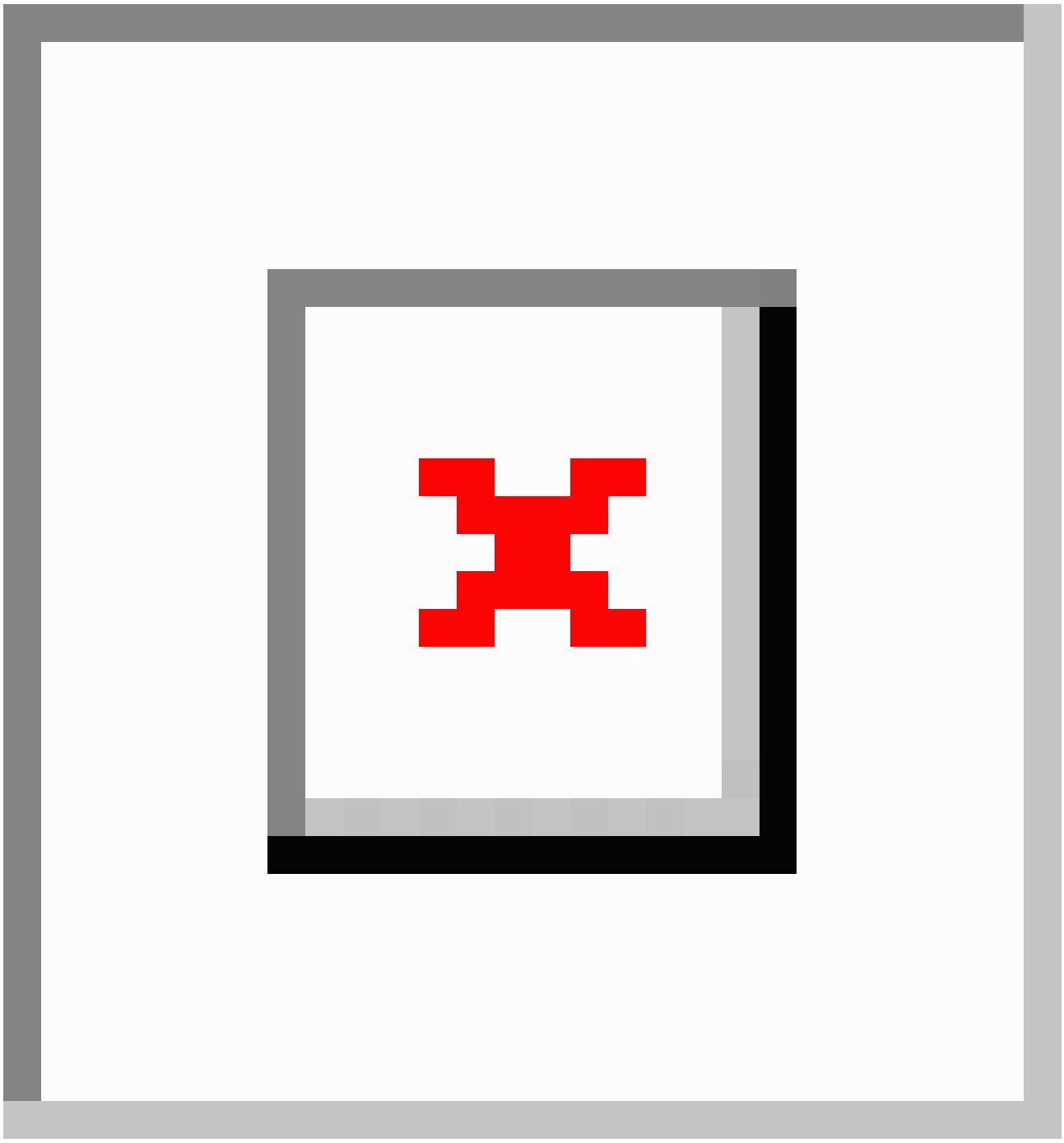


Figure 4. Tolvaptan Monitoring Summary. eGFR: estimated glomerular filtration rate.

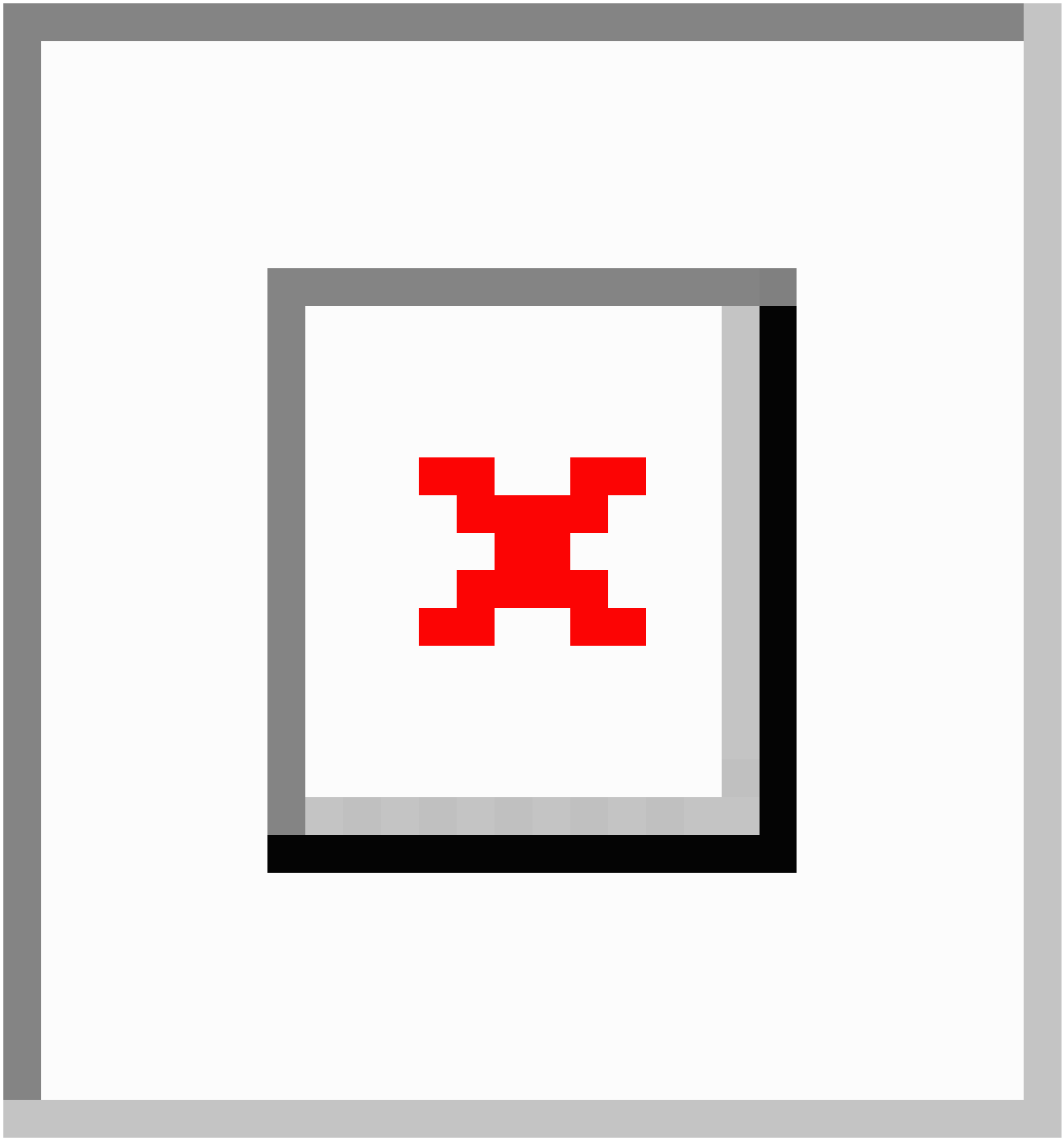


Figure 5. Tolvaptan Monitoring Metrics. ALK Phos: alkaline phosphatase; ALT: alanine transaminase; AST: aspartate aminotransferase; eGFR: estimated glomerular filtration rate; Tot Bili: total bilirubin.

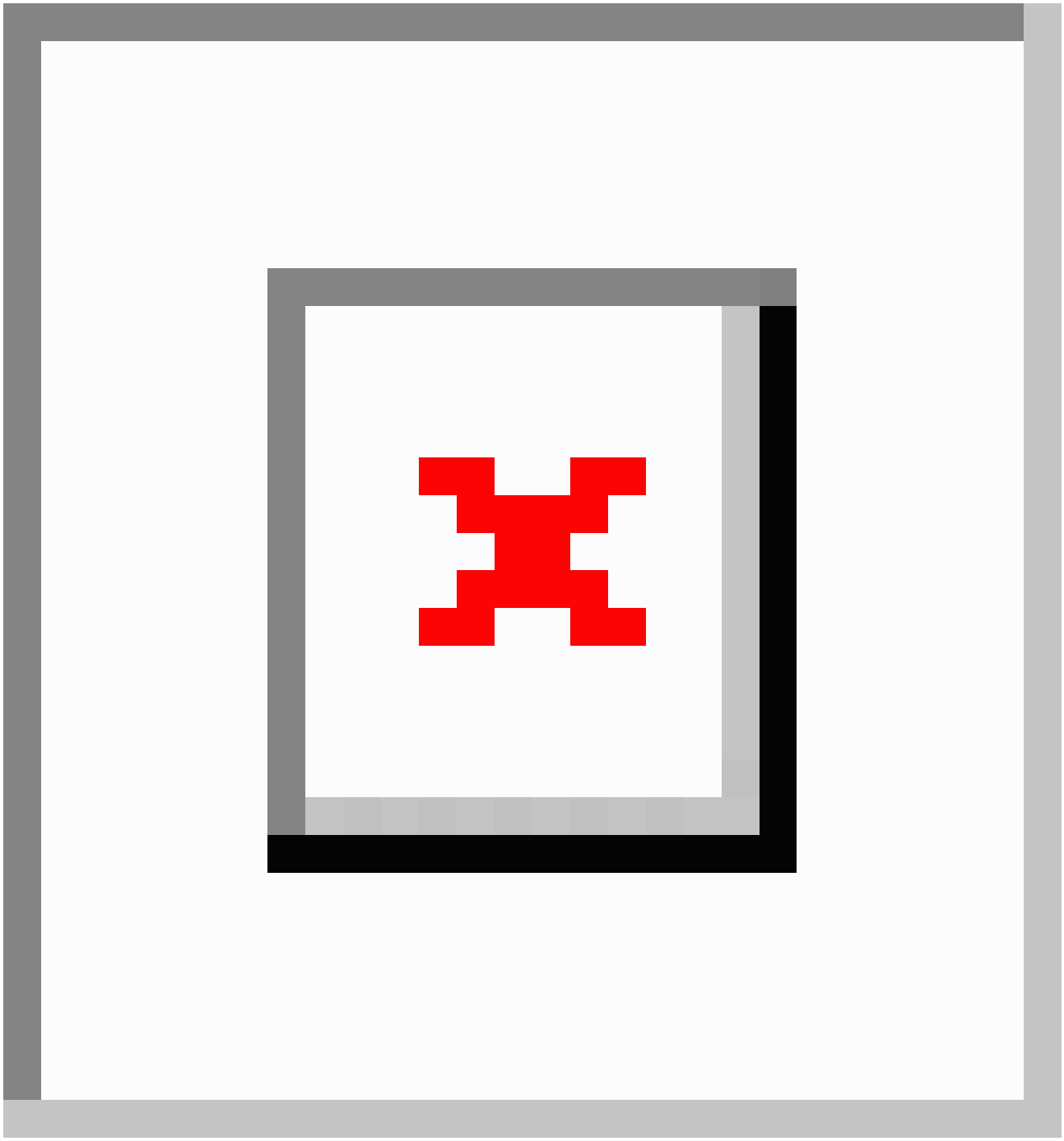
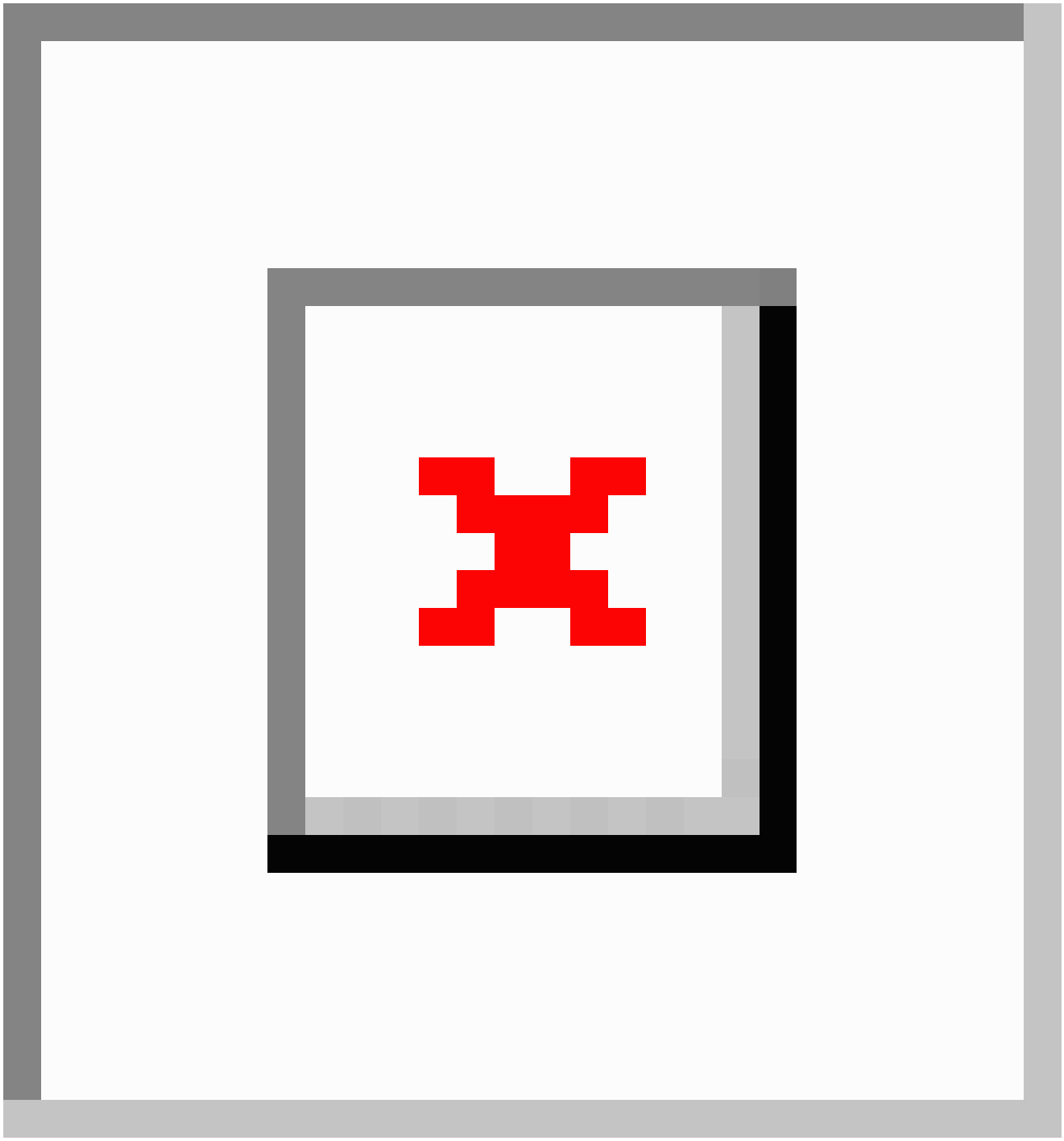


Figure 6. Tolvaptan detailed summary report.



Characteristics of Enrolled Patients

As of September 1, 2022, a total of 214 patients have been enrolled in the tolvaptan program across the Mayo Clinic Health System in Minnesota, Arizona, Florida, and Wisconsin (Table 1).

Of the 214 patients, 126 (59%) were enrolled in the Tolvaptan Monitoring Registry, and the remaining 88 patients were included in Past Tolvaptan Patients. A total of 10 nephrologists were caring for these patients across the 4 locations. Table 2 displays characteristics of the patients in the Tolvaptan Monitoring Registry, including their demographics and MIC status.

The registry included 57.9% (n=124) female, 96.2% (n=206) White, 2.8% (n=6) Hispanic, and 0.9% (n=2) African American individuals. The mean age at enrollment was 43.1 (SD 9.9) years, and 86 patients had a documented MIC. Most patients had an MIC of 1C (n=38, 44.2%), followed by 1D (n=25, 29.1%) and 1E (n=19, 22.1%). Additionally, 3 patients (5.1%) had an MIC of 1B but were prescribed tolvaptan based on a non-MIC criterion. Of note, 33 (15.4%) patients were taking tolvaptan as part of a clinical trial and remained on the drug after trial completion (following FDA approval of the drug). These patients were added retrospectively to the application because they were taking tolvaptan prior to the creation of the EHR-based application.

Table . Distribution of patients with autosomal dominant polycystic kidney disease receiving tolvaptan treatment across the Mayo Clinic system.

Region	Tolvaptan Monitoring Registry (n=126), n (%)	Past Tolvaptan Patients (n=88), n (%)	Total (N=214), n (%)
Minnesota	80 (63.5)	75 (85.2)	155 (72.4)
Arizona	17 (13.5)	7 (8)	24 (11.2)
Florida	22 (17.4)	4 (4.5)	26 (12.1)
MCHS[a], WI[b]	7 (5.6)	2 (2.3)	9 (4.2)

[a]MCHS: Mayo Clinic Health System.

[b]WI: Wisconsin.

Table . Patient demographics and Mayo imaging class.

		Tolvaptan Monitoring Registry	Past Tolvaptan Patients	Total patients
Demographics				
	Patients, n	126	88	214
	Female, n (%)	72 (57.1)	52 (59)	124 (58)
	Age at registry inclusion (years), mean (SD)	43.8 (9.9)	40.4 (9.9)	43.1 (9.9)
	White, n (%)	123 (97.6)	83 (94.3)	206 (96.3)
	Hispanic, n (%)	3 (2.4)	3 (3.4)	6 (2.8)
	African American, n (%)	0 (0)	2 (2.3)	2 (0.9)
	Enrollment through clinical trials, n (%)	29 (22.4)	4 (4.5)	33 (15.4)
Mayo imaging class				
	Patients, n	58	28	86
	1A	0 (0)	0 (0)	0 (0)
	1B	3 (5.2)	1 (3.6)	4 (4.7)
	1C	27 (46.6)	11 (39.3)	38 (44.2)
	1D	19 (32.8)	6 (21.4)	25 (29.1)
	1E	9 (15.5)	10 (35.7)	19 (22.1)

Outcomes of Using the Tolvaptan Application

The implementation of the tolvaptan EHR-based application streamlined the monitoring process of patients treated with tolvaptan in several ways. First, the automated addition of patients into the registry reduced the risk of missing any patients started on tolvaptan, thus assuring that all treated patients were closely monitored for any adverse events that might occur while on therapy. Second, the application allowed for efficient and timely identification of patients who had abnormal laboratory test results and enabled nursing outreach to patients who might need further intervention or education on medication management. Overall, 20 (9.3%) patients had liver function test abnormalities, but only 5 (2.3%) had to discontinue the drug because of hepatotoxicity. The most common reason for drug discontinuation was related to the aquaretic effect, in 10 patients (4.7%), while only 4 (1.8%) could not continue in the program because of medical insurance–related issue. Third, the application provides a comprehensive and up-to-date summary of all pertinent clinical information related to the management of ADPKD, including medications, appointments, laboratory tests, and notes from the care team. Fourth, the application facilitated communication and collaboration among the multidisciplinary team involved in the care of patients with ADPKD. The standardization process and easy data access to all enrolled patients in the registry provided an opportunity for the care team to meet quarterly to review and discuss specific scenarios. These discussions sometimes led team members to share their experiences regarding challenging situations or drug-related adverse events or drug intolerance and at other times to propose enhancements to the EHR application. Finally, the application enhanced the efficiency of the tolvaptan program by reducing the time and effort (informally reported by the care team) required for enrollment, tracking, and monitoring of patients. More specifically, for the physicians, the only required task to enroll a patient in the program was identifying the candidate and making an electronic referral to the nursing team. The nursing team then initiated the education, treatment, and monitoring without any further escalation to the physician, unless there were concerns. For the nurses, all relevant information was consolidated, reducing the need to navigate to various EHR screens and modules.

Discussion

Principal Findings

In this report, we share our experience in developing and implementing an EHR application to manage and monitor patients with ADPKD who were enrolled in the tolvaptan program across several sites at our institution. This application streamlined the clinical workflow and enabled the nephrology nursing team to proactively take appropriate action to mitigate drug-related serious adverse events. Tolvaptan is the first and only FDA-approved drug to slow the progression of ADPKD, but it has multiple adverse effects, most seriously liver toxicity, which can be potentially severe, albeit rare. Therefore, frequent laboratory test monitoring is required to detect early drug-related toxicity. This application is crucial in facilitating the monitoring of patients taking tolvaptan, especially in large centers with high case load or smaller clinics with limited staff and resources. Overall, 20 (9.3%) patients had liver function test abnormalities, but only 5 (2.3%) had to discontinue the drug because of hepatotoxicity. The frequency of these events is very similar to those reported in the REPRISE clinical trial (10.9% hepatotoxicity and 1.6% discontinuation for a liver event) [6]. The most common reason for drug discontinuation was related to the aquaretic effect, which occurred in 4.7% (n=10) of patients. This is higher than the frequency reported in the REPRISE trial (2.1%) but not surprising since participants enrolled in clinical trials may be more motivated to adhere to the treatment protocol. Nonetheless, these clinical outcomes are reassuring.

The logistical requirements of any tolvaptan program may limit the ability of nephrology practices to provide this effective therapy. With the shortage of physicians and their high level of burnout [17-19], well-designed EHR integration that helps review, in a consolidated manner, relevant data for clinical care is important, although it comes with a higher up-front cost [20-23]. This is now more relevant because about 90% of office-based physicians in the United States use an EHR [24], and higher perceived EHR usability is associated with higher levels of perceived positive outcomes (improved patient care) and lower levels of perceived negative outcomes (worse patient interactions and work-life integration) [25]. Whether developing such digital systems is worth the investment is a relevant question for health care systems [26], but they can certainly be scaled in real-world settings. The cost of creating similar EHR-based applications will vary depending on each organization structure and is mostly an up-front cost. This includes the time required by both the clinical care team (nurses, physicians, and other clinical staff) and informatics team (program manager and technical build team) to identify the clinical need and develop and test the product. For our practice, it required at least 1 physician and 1 nurse champion to be present at each meeting (4 staff members were engaged) with the informatics team for 1 hour every 2 weeks on average over a 6-month period (12 hours per staff member involved). Since all our staff are salary based, this work was primarily supported by discretionary efforts and during nonclinical activities (lunch hour or administrative time). Regarding the informatics team (CKIS), our institution has allocated an operational budget to support various EHR-related projects across the enterprise; thus, we did not have to request extra funds to support this effort.

The advantages of these applications and data analytics capabilities within the EHR have been well described for various diseases and conditions, recently including more COVID-19–related activities, to manage the clinical practice safely [27-37]. Besides keeping track of a defined patient population, aggregating data, and identifying care gaps, communication with patients through the patient portal is readily accessible. In addition, bulk messaging (sending the same message to a group of patients in one click) is a convenient feature of the application. For example, staff can easily remind patients to do their monthly or quarterly laboratory tests when these have lapsed and do a synchronous or asynchronous quick health check if needed.

Limitations

The design, development, and deployment in clinical practice of this integrated digital application has limitations. The process is iterative and requires buy-in from various stakeholders, an up-front investment in time, resources, and change management capabilities. Our clinical team was receptive, open to change, and willing to embrace new workflows because of the perceived value of adopting the application (more efficient and safer care delivery). One limitation of our study is that it was conducted in a single health care system. However, the successful implementation of this application in our Minnesota practice, followed by its expansion to all Mayo Clinic practices, highlights the potential for scaling to other health care systems. Another limitation of our study is the lack of objective efficiency outcome measures. The workflow improvement and satisfaction were not evalutated in a formal manner by the physicians and nurses. Ideally, our study would assess the impact of the application using (1) direct observation (time-motion studies), (2) EHR log-based analysis (EHR log data), (3) care team pre- and postimplementation surveys, or a combination of these. However, prior to the implementation of the application, the management of the tolvaptan program was ad hoc, carried out by a care team that performed multiple unrelated clinical activities. This made it impractical to use time-motion studies and impossible to meaningfully use EHR log data. A care team survey was considered, but because the transition to the application was done during the COVID-19 pandemic when our personnel resources were very strained, noncritical activities were paused. Prospective studies are necessary to validate the effectiveness of this application and its potential for improving care processes and ultimately patient outcomes.

Conclusion

In conclusion, our multidisciplinary team developed an EHR-integrated digital monitoring protocol that could facilitate safe, efficient, and high-quality care for patients with ADPKD who were prescribed tolvaptan. The implementation of this application in our health care system can be scaled to other health care systems or smaller clinics after further validation. This can reduce some barriers and help safely provide the best available treatment for eligible patients.

Conflicts of Interest

ZZ serves as a member of the Epic nephrology steering board committee. MH has received consulting fees from Otsuka in the past for work unrelated to this study. All other authors report no conflicts of interest.

Multimedia Appendix 1
Definitions of terms in columns and flags.
[DOCX File, 13 KB - medinform_v12i1e50164_app1.docx]

References

1. Shukoor SS, Vaughan LE, Edwards ME, et al. Characteristics of patients with end-stage kidney disease in ADPKD. Kidney Int Rep 2020 Dec;6(3):755-767. [doi: 10.1016/j.ekir.2020.12.016] [Medline: 33732990]
2. Hateboer N, v Dijk MA, Bogdanova N, et al. Comparison of phenotypes of polycystic kidney disease types 1 and 2. Lancet 1999 Jan;353(9147):103-107. [doi: 10.1016/S0140-6736(98)03495-3] [Medline: 10023895]
3. Chebib FT, Torres VE. Autosomal dominant polycystic kidney disease: core curriculum 2016. Am J Kidney Dis 2016 May;67(5):792-810. [doi: 10.1053/j.ajkd.2015.07.037] [Medline: 26530876]
4. Irazabal MV, Rangel LJ, Bergstralh EJ, et al. Imaging classification of autosomal dominant polycystic kidney disease: a simple model for selecting patients for clinical trials. J Am Soc Nephrol 2015 Jan;26(1):160-172. [doi: 10.1681/ASN.2013101138] [Medline: 24904092]
5. Torres VE, Chapman AB, Devuyst O, et al. Tolvaptan in patients with autosomal dominant polycystic kidney disease. N Engl J Med 2012 Dec 20;367(25):2407-2418. [doi: 10.1056/NEJMoa1205511] [Medline: 23121377]
6. Torres VE, Chapman AB, Devuyst O, et al. Tolvaptan in later-stage autosomal dominant polycystic kidney disease. N Engl J Med 2017 Nov 16;377(20):1930-1942. [doi: 10.1056/NEJMoa1710030] [Medline: 29105594]
7. Tai-Seale M, Olson CW, Li J, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff (Millwood) 2017 Apr 1;36(4):655-662. [doi: 10.1377/hlthaff.2016.0811] [Medline: 28373331]
8. Prasad K, Poplau S, Brown R, et al. Time pressure during primary care office visits: a prospective evaluation of data from the Healthy Work Place study. J Gen Intern Med 2020 Feb;35(2):465-472. [doi: 10.1007/s11606-019-05343-6] [Medline: 31797160]
9. Witry M, Marie BS, Reist J. Provider perspectives and experiences following the integration of the prescription drug monitoring program into the electronic health record. Health Informatics J 2022;28(3):14604582221113435. [doi: 10.1177/14604582221113435] [Medline: 35829729]
10. Seki T, Aki M, Furukawa TA, et al. Electronic health record-nested reminders for serum lithium level monitoring in patients with mood disorder: randomized controlled trial. J Med Internet Res 2023 Mar 22;25:e40595. [doi: 10.2196/40595] [Medline: 36947138]
11. Mishra V, Chouinard M, Keiser J, et al. Automating vancomycin monitoring to improve patient safety. Jt Comm J Qual Patient Saf 2019 Nov;45(11):757-762. [doi: 10.1016/j.jcjq.2019.07.001] [Medline: 31526711]
12. Bundy DG, Marsteller JA, Wu AW, et al. Electronic health record-based monitoring of primary care patients at risk of medication-related toxicity. Jt Comm J Qual Patient Saf 2012 May;38(5):216-223. [doi: 10.1016/s1553-7250(12)38027-6] [Medline: 22649861]
13. Neprash HT, Vock DM, Hanson A, et al. Effect of integrating access to a prescription drug monitoring program within the electronic health record on the frequency of queries by primary care clinicians: a cluster randomized clinical trial. JAMA Health Forum 2022 Jun;3(6):e221852. [doi: 10.1001/jamahealthforum.2022.1852] [Medline: 35977248]
14. Lee SY, Cherian R, Ly I, Horton C, Salley AL, Sarkar U. Designing and implementing an electronic patient registry to improve warfarin monitoring in the ambulatory setting. Jt Comm J Qual Patient Saf 2017 Jul;43(7):353-360. [doi: 10.1016/j.jcjq.2017.03.006] [Medline: 28648221]
15. Chebib FT, Torres VE. Assessing risk of rapid progression in autosomal dominant polycystic kidney disease and special considerations for disease-modifying therapy. Am J Kidney Dis 2021 Aug;78(2):282-292. [doi: 10.1053/j.ajkd.2020.12.020] [Medline: 33705818]
16. Chebib FT, Perrone RD, Chapman AB, et al. A practical guide for treatment of rapidly progressive ADPKD with tolvaptan. J Am Soc Nephrol 2018 Oct;29(10):2458-2470. [doi: 10.1681/ASN.2018060590] [Medline: 30228150]
17. Physician workforce projections: the complexities of physician supply and demand. Association of American Medical Colleges. 2021. URL: https://www.aamc.org/data-reports/workforce/report/physician-workforce-projections [accessed 2024-04-23]
18. Shanafelt TD, West CP, Dyrbye LN, et al. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. Mayo Clin Proc 2022 Dec;97(12):2248-2258. [doi: 10.1016/j.mayocp.2022.09.002] [Medline: 36229269]

19. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 1;27(4):531-538. [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
20. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426. [doi: 10.1370/afm.2121] [Medline: 28893811]
21. Kroth PJ, Morioka-Douglas N, Veres S, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 2;2(8):e199609. [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
22. Tawfik DS, Sinha A, Bayati M, et al. Frustration with technology and its relation to emotional exhaustion among health care workers: cross-sectional observational study. J Med Internet Res 2021 Jul 6;23(7):e26817. [doi: 10.2196/26817] [Medline: 34255674]
23. Sinsky CA, Shanafelt TD, Ripp JA. The electronic health record inbox: recommendations for relief. J Gen Intern Med 2022 Nov;37(15):4002-4003. [doi: 10.1007/s11606-022-07766-0] [Medline: 36036837]
24. Electronic medical records/electronic health records (EMRs/EHRs). US Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm [accessed 2023-01-31]
25. Melnick ER, Sinsky CA, Dyrbye LN, et al. Association of perceived electronic health record usability with patient interactions and work-life integration among US physicians. JAMA Netw Open 2020 Jun 1;3(6):e207374. [doi: 10.1001/jamanetworkopen.2020.7374] [Medline: 32568397]
26. Shanafelt TD, Larson D, Bohman B, et al. Organization-wide approaches to foster effective unit-level efforts to improve clinician well-being. Mayo Clin Proc 2023 Jan;98(1):163-180. [doi: 10.1016/j.mayocp.2022.10.031] [Medline: 36603944]
27. Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752. [doi: 10.7326/0003-4819-144-10-200605160-00125] [Medline: 16702590]
28. Dreyer NA, Garner S. Registries for robust evidence. JAMA 2009 Aug 19;302(7):790-791. [doi: 10.1001/jama.2009.1092] [Medline: 19690313]
29. Hersh W. Electronic health records facilitate development of disease registries and more. Clin J Am Soc Nephrol 2011 Jan;6(1):5-6. [doi: 10.2215/CJN.09901110] [Medline: 21127135]
30. Jaffe MG, Lee GA, Young JD, Sidney S, Go AS. Improved blood pressure control associated with a large-scale hypertension program. JAMA 2013 Aug 21;310(7):699-705. [doi: 10.1001/jama.2013.108769] [Medline: 23989679]
31. Jolly SE, Navaneethan SD, Schold JD, et al. Development of a chronic kidney disease patient navigator program. BMC Nephrol 2015 May 3;16:69. [doi: 10.1186/s12882-015-0060-2] [Medline: 26024966]
32. Mendu ML, Waikar SS, Rao SK. Kidney disease population health management in the era of accountable care: a conceptual framework for optimizing care across the CKD spectrum. Am J Kidney Dis 2017 Jul;70(1):122-131. [doi: 10.1053/j.ajkd.2016.11.013] [Medline: 28132720]
33. Navaneethan SD, Jolly SE, Schold JD, et al. Pragmatic randomized, controlled trial of patient navigators and enhanced personal health records in CKD. Clin J Am Soc Nephrol 2017 Sep 7;12(9):1418-1427. [doi: 10.2215/CJN.02100217] [Medline: 28778854]
34. Rana JS, Karter AJ, Liu JY, Moffet HH, Jaffe MG. Improved cardiovascular risk factors control associated with a large-scale population management program among diabetes patients. Am J Med 2018 Jun;131(6):661-668. [doi: 10.1016/j.amjmed.2018.01.024] [Medline: 29576192]
35. Mendu ML, Ahmed S, Maron JK, et al. Development of an electronic health record-based chronic kidney disease registry to promote population health management. BMC Nephrol 2019 Mar 1;20(1):72. [doi: 10.1186/s12882-019-1260-y] [Medline: 30823871]
36. Peralta CA, Livaudais-Toman J, Stebbins M, et al. Electronic decision support for management of CKD in primary care: a pragmatic randomized trial. Am J Kidney Dis 2020 Nov;76(5):636-644. [doi: 10.1053/j.ajkd.2020.05.013] [Medline: 32682696]
37. Jose T, Warner DO, O'Horo JC, et al. Digital health surveillance strategies for management of coronavirus disease 2019. Mayo Clin Proc Innov Qual Outcomes 2021 Feb;5(1):109-117. [doi: 10.1016/j.mayocpiqo.2020.12.004] [Medline: 33521582]

Abbreviations

ADPKD: autosomal dominant polycystic kidney disease
ALT: alanine transaminase
AST: aspartate transaminase
CKIS: Cohort Knowledge Intelligent Solutions
CYP3A: cytochrome P450, family 3, subfamily A
eGFR: estimated glomerular filtration rate

EHR: electronic health record
ESKF: end-stage kidney failure
FDA: US Food and Drug Administration
MIC: Mayo imaging classification
PDMP: prescription drug monitoring program
PKD: polycystic kidney disease
REMS: risk evaluation and mitigation strategy
TKV: total kidney volume

Edited by C Perrin; submitted 21.06.23; peer-reviewed by C Kwoh, J Walsh, O Amro; revised version received 06.03.24; accepted 25.03.24; published 01.05.24.

Please cite as:
Chedid M, Chebib FT, Dahlen E, Mueller T, Schnell T, Gay M, Hommos M, Swaminathan S, Garg A, Mao M, Amberg B, Balderes K, Johnson KF, Bishop A, Vaughn JK, Hogan M, Torres V, Chaudhry R, Zoghby Z
An Electronic Health Record–Integrated Application for Standardizing Care and Monitoring Patients With Autosomal Dominant Polycystic Kidney Disease Enrolled in a Tolvaptan Clinic: Design and Implementation Study
JMIR Med Inform 2024;12:e50164
URL: https://medinform.jmir.org/2024/1/e50164
doi:10.2196/50164

Original Paper

The Implementation of an Electronic Medical Record in a German Hospital and the Change in Completeness of Documentation: Longitudinal Document Analysis

Florian Wurster[1], MSc; Marina Beckmann[1], DPhil; Natalia Cecon-Stabel[1], MSc; Kerstin Dittmer[1], MA; Till Jes Hansen[1], MSc; Julia Jaschke[2], MSc; Juliane Köberlein-Neu[2], Prof Dr; Mi-Ran Okumu[1], MA; Carsten Rusniok[1], MA; Holger Pfaff[1], Prof Dr; Ute Karbach[1], PD, Dr

[1]Chair of Quality Development and Evaluation in Rehabilitation, Institute of Medical Sociology, Health Services Research, and Rehabilitation Science, Faculty of Human Sciences & Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany

[2]Center for Health Economics and Health Services Research, University of Wuppertal, Wuppertal, Germany

Corresponding Author:
Florian Wurster, MSc
Chair of Quality Development and Evaluation in Rehabilitation, Institute of Medical Sociology, Health Services Research, and Rehabilitation Science, Faculty of Human Sciences & Faculty of Medicine and University Hospital Cologne, University of Cologne
Eupener Str. 129
Cologne, 50933
Germany
Phone: 49 22147897116
Email: florian.wurster@uni-koeln.de

Abstract

Background: Electronic medical records (EMR) are considered a key component of the health care system's digital transformation. The implementation of an EMR promises various improvements, for example, in the availability of information, coordination of care, or patient safety, and is required for big data analytics. To ensure those possibilities, the included documentation must be of high quality. In this matter, the most frequently described dimension of data quality is the completeness of documentation. In this regard, little is known about how and why the completeness of documentation might change after the implementation of an EMR.

Objective: This study aims to compare the completeness of documentation in paper-based medical records and EMRs and to discuss the possible impact of an EMR on the completeness of documentation.

Methods: A retrospective document analysis was conducted, comparing the completeness of paper-based medical records and EMRs. Data were collected before and after the implementation of an EMR on an orthopaedical ward in a German academic teaching hospital. The anonymized records represent all treated patients for a 3-week period each. Unpaired, 2-tailed *t* tests, chi-square tests, and relative risks were calculated to analyze and compare the mean completeness of the 2 record types in general and of 10 specific items in detail (blood pressure, body temperature, diagnosis, diet, excretions, height, pain, pulse, reanimation status, and weight). For this purpose, each of the 10 items received a dichotomous score of 1 if it was documented on the first day of patient care on the ward; otherwise, it was scored as 0.

Results: The analysis consisted of 180 medical records. The average completeness was 6.25 (SD 2.15) out of 10 in the paper-based medical record, significantly rising to an average of 7.13 (SD 2.01) in the EMR (t_{178}=–2.469; *P*=.01; *d*=–0.428). When looking at the significant changes of the 10 items in detail, the documentation of diet (*P*<.001), height (*P*<.001), and weight (*P*<.001) was more complete in the EMR, while the documentation of diagnosis (*P*<.001), excretions (*P*=.02), and pain (*P*=.008) was less complete in the EMR. The completeness remained unchanged for the documentation of pulse (*P*=.28), blood pressure (*P*=.47), body temperature (*P*=.497), and reanimation status (*P*=.73).

Conclusions: Implementing EMRs can influence the completeness of documentation, with a possible change in both increased and decreased completeness. However, the mechanisms that determine those changes are often neglected. There are mechanisms that might facilitate an improved completeness of documentation and could decrease or increase the staff's burden caused by

documentation tasks. Research is needed to take advantage of these mechanisms and use them for mutual profit in the interests of all stakeholders.

Trial Registration: German Clinical Trials Register DRKS00023343; https://drks.de/search/de/trial/DRKS00023343

(JMIR Med Inform 2024;12:e47761) doi:10.2196/47761

KEYWORDS

clinical documentation; digital transformation; document analysis; electronic medical record; EMR; Germany; health services research; hospital; implementation

Introduction

The digital transformation of the health care system is considered an essential subject to meet current and future societal challenges such as an aging population or rising health care expenditures while at the same time maintaining a high quality of care [1]. An important early step in hospitals' digitalization and a fundamental requirement for expanding digital maturity is the implementation of an electronic medical record (EMR) [2]. This EMR is considered to be an **"**electronic record of health care information of an individual that is created, gathered, managed, and consulted by authorized clinicians and staff within 1 health care organization" [3] and replaces the internal clinical documentation on preprinted paper-based charts. Studies show that the implementation of an EMR can lead to various improvements in the clinical context (eg, in the availability of information [4], coordination of care [5], or patient safety [6]). Moreover, the EMR facilitates the secondary usage of the documented data for research purposes through its digital accessibility [7]. To reach those benefits, it is indispensable that the EMR contain documentation that is of high quality. However, there are varying definitions regarding the quality of documentation. In that matter, the Institute of Medicine defined completeness, legibility, accuracy, and meaning as the main aspects of a medical record's data quality [8]. For those, the completeness of documentation was shown to be the most common dimension of data quality when empirically analyzing the documentation in EMRs [9], and it was highlighted to be especially important for secondary uses such as big data analyses [10].

Our recent systematic review also stated the completeness of documentation as the state of the art for the comparison of paper-based and EMRs [11]. This comparison is important since the implementation of an EMR and the associated transition from handwritten documentation to digital documentation can heavily affect the documentation subject since the transition offers the possibility to adjust which information has to be documented in which way [12]. For example, digitization enables the adoption of certain functionalities that can alter the completeness of documentation, like automatically transferring information from other digital devices to the EMR [13]. Moreover, when working with the EMR, information can be documented remotely, while the paper-based medical record had to be located and physically accessed first. In this matter, several studies conducted in the inpatient setting showed increased completeness in the EMR compared to the paper-based medical record, for example, for the documentation of signs and symptoms [13,14], weight and height, or malnutrition screening [15]. This suggests that the implementation of an EMR might lead to improvements in the completeness of documentation in general. It is therefore the main purpose of this study to evaluate the change in completeness due to the implementation of an EMR in an inpatient setting. Literature already provides proof of a change of completeness in regard to some specific documented information that is analyzed in this work (eg, the documentation of vital signs) [13,14]. Those empirical results might thus be validated for the presented work's specific setting and discipline. In addition, some of the information that is analyzed in this work is not described in literature yet (eg, the documentation of pain). It is examined for the first time with regard to changes in completeness after the implementation of an EMR.

The knowledge gained can not only support the implementation of new EMRs but could also help understand and optimize arising changes in documentation when existing EMRs need to be adapted [16,17]. This is an important aspect, as the implementation of new EMRs is described as one of the most important interventions to improve the quality of documentation [18]. In this process, mechanisms affecting the completeness of documentation in medical records are not completely understood [10]. On the other hand, this knowledge is needed to fulfill reported educational needs regarding how to reach the optimum quality of documentation [19]. In this context, this study contributes to a more comprehensive understanding of the impact of an EMR on the quality of documentation.

Methods

Overview

This study follows the "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE) statement [20] whenever it is applicable. It offers reporting standards to ensure the reporting of any important information in empirical research studies. A checklist with details, where the STROBE information is mentioned in the manuscript, can be found in Multimedia Appendix 1.

Ethical Considerations

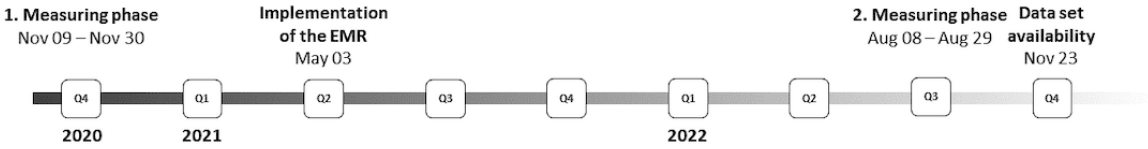
The study has been approved by the ethics committee of the Medical Faculty of the University of Cologne, Germany (20-1349). All data was anonymized at all times during the scientific analysis. No compensation was paid.

Setting and Participants

The study took place as part of the research project eCoCo, which Beckmann et al [21] described in detail. Within the eCoCo project, the researchers collected various types of data

(observations, surveys, interviews, documents, and administrative data) to investigate a possible change in interprofessional collaboration and clinical workflows following the implementation of an inpatient EMR. This study is part of the related work package on documentation content and quality, which took place in a large academic teaching hospital in Germany. The hospital replaces its internal documentation on preprinted paper-based charts with a commercial EMR system (Meona; Mesalvo Freiburg GmbH). The EMR runs on multiple computers that can be moved flexibly over the ward on trolleys. The study follows a pre-post design, retrospectively analyzing the content of the medical records before and after the implementation of the EMR on the hospitals' orthopaedical ward. Within the first measuring phase, the paper-based medical records were provided as a digital copy of the paper sheets. Those paper-based records represent all patients who were treated on the ward during the last 3 weeks in November 2020 (t0). After 6 months, employees received training on how to use the EMR before the implementation of the EMR took place in May 2021. The EMRs were again provided as a digital copy within a second measuring phase, representing all patients who were treated on the same ward during the first 3 weeks of August 2022 (t1). This resulted in a gap of 15 months between the first and second measuring phases. The complete data set was available to the research team in November 2022 (Figure 1). The hospital provided anonymized medical records to the research team after the records were archived and cleared of sensitive personal data (eg, the patient's name or date of birth) in the hospital's internal processes. Any assignment of the patient data or linking of the records' contents to any individual patient was therefore impossible for the research team, which is, thus, in compliance with the European Union General Data Protection Regulation. This also implies the absence of sociodemographic information for describing the compared samples. The hospital's mandatory annual quality report, which is available to the public through a designated database [22], is therefore used to describe the ward's patient sample and the performed treatments in general. This allows an approximation to a description and comparison of the compared samples in terms of their *International Classification of Diseases* (*ICD*)–diagnoses distribution.

Figure 1. Data collection. EMR: electronic medical record.

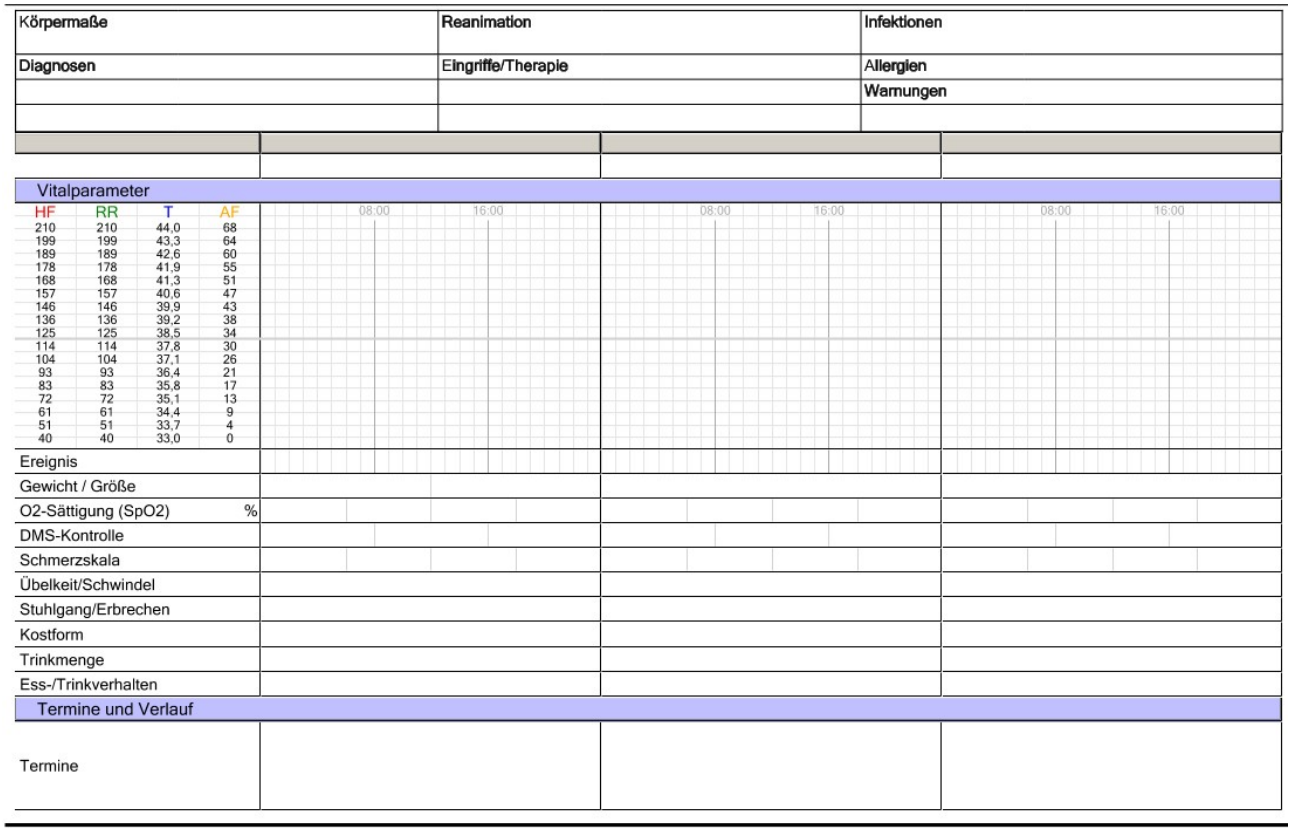


Study Objective

To answer the question of a possible change in completeness, the records were analyzed by content [23]. The change from paper-based documentation to EMRs always offers the opportunity to fundamentally change the structure of the records. This was shown exemplarily by Montagna et al [12] when the documentation as a continuously written text in the paper-based record was changed to a list of events in the EMR. It is therefore important to ensure the comparability between the 2 record types for the purpose of analyzing a possible change in their completeness. To achieve comparability, the medical records progress note was selected as a specific object of interest for this study's analyses since it retained the same structure and format in both record types. Part of this progress note is the fever chart (Figures 2 and 3), which includes basic details about vital signs, personal health data, etc [24].

Figure 2. Paper-based fever chart.

Figure 3. Electronic fever chart.



All information that was commonly documented in both of the 2 record types (paper-based and electronic) became part of this work. Weiskopf and Weng [9] described this selection mechanism for assessing data quality based on the parallels between the EMR and the paper-based record. This procedure resulted in a total of 10 key items that were analyzed for completeness in this work: blood pressure, body temperature, diagnosis, diet, excretions, height, pain, pulse, reanimation

status, and weight. The documentation of this information is equally possible and performed by nurses and physicians. However, there is no information available about who specifically entered the information.

All of those items should be documented immediately when patient care begins on the ward [25]. However, while the documentation of vital signs can take place up to several times a day, the documentation of the patient's diet usually occurs once a day, and the documentation of the reanimation status (patient's preference regarding a possible resuscitation) is probably documented only once per hospital stay. Because of these varying documentation practices and to ensure comparability, the analysis focuses on certain documentation in the progress notes that was entered on the first day of patient care on the ward. With regard to the documentation of a diagnosis, it is therefore the diagnosis with which a patient is admitted to the hospital. This diagnosis is mainly responsible for the allocation to specific medical specialties as well as a certain ward and does not necessarily have to match the final diagnosis at the time of discharge, which is important for reimbursement purposes.

Statistical Analysis of Completeness

For every record, each of the 10 items received a dichotomous score of 1 if it was documented on the first day of patient care on the ward; otherwise, it was scored as 0. This resulted in a percentage of completeness for each item per record type. Chi-square tests for independence were used to assess statistically significant differences in the percentage of completeness per item between the 2 record types. Relative risks were calculated for the association between the electronic record type and a possible increase in completeness. To improve the reliability of the associated confidence intervals, they were calculated with 5000 bootstrap replications since the original sample sizes are unbalanced. Moreover, the overall completeness was assessed as sum of the 10 items, resulting in a mean score of completeness per record type ranging from 0 (no item documented) to 10 (all 10 items documented). Those mean scores of completeness per record type were analyzed for equality of variance and statistical difference using unpaired, 2-tailed *t* tests. Assumptions were checked using several methods (normal distribution: QQ plots and Shapiro-Wilk test; homogeneity of variances: Levene test; and linearity: scatter plot). The level of significance was set to be $P<.05$ for all calculations. The data were stored in Microsoft Excel (Microsoft Corp) and analyzed in December 2022 using SPSS software (version 29; IBM Corp).

Results

Participants

During the first measuring phase (November 2020), a total of 44 patients (paper-based) were treated on the orthopaedical ward. They were encountering a total of 136 treated patients (electronic) during the second measuring phase (August 2022). This resulted in a total of 180 medical records that became part of this analysis. Due to the data protection regulation and the accompanied anonymization of the records data, there is no information regarding the demographics of the specific study population. Therefore, the ward's ICD-diagnosis distribution is given as an approximation of a sample description. In 2020, the 3 most frequently coded diagnoses for the orthopaedical ward were complications of internal orthopedic prosthetic devices, implants and grafts (ICD-T84), dorsalgia (ICD-M54), and fracture of shoulder and upper arm (ICD-S42). This report is not yet published for 2022, but the top 3 treated diagnoses in 2019 or 2021 were similar to those in 2020 (Table 1). It can therefore be expected that the treated diagnoses will be similar in 2022, too. Another supporting fact is that the most frequently performed procedure (surgical access to the lumbar spine, the sacrum, or the coccyx [coded as OPS-5-032 in the German adaptation of the International Classification of Procedures in Medicine which is part of the coding system for hospitals reimbursement]) was the same in all 3 years (2019-2021).

Table 1. Most frequently coded diagnoses.

ICD[a] Code	Values, n[b]/N[c] (%)
2019	
Dorsalgia (ICD-M54)	213/3147 (6.77)
Other spondylopathies (ICD-M48)	133/3147 (4.23)
Fracture of forearm (ICD-S52)	131/3147 (4.16)
2020	
Complications of internal orthopedic prosthetic devices, implants and grafts (ICD-T84)	166/2912 (5.7)
Dorsalgia (ICD-M54)	148/2912 (5.08)
Fracture of shoulder and upper arm (ICD-S42)	121/2912 (4.16)
2021	
Complications of internal orthopedic prosthetic devices, implants and grafts (ICD-T84)	164/3091 (5.3)
Dorsalgia (ICD-M54)	163/3091 (5.27)
Fracture of forearm (ICD-S52)	159/3091 (5.14)

[a]ICD: International Classification of Diseases.

[b]Frequency of coded diagnosis.

[c]Total inpatient cases.

Change of Completeness

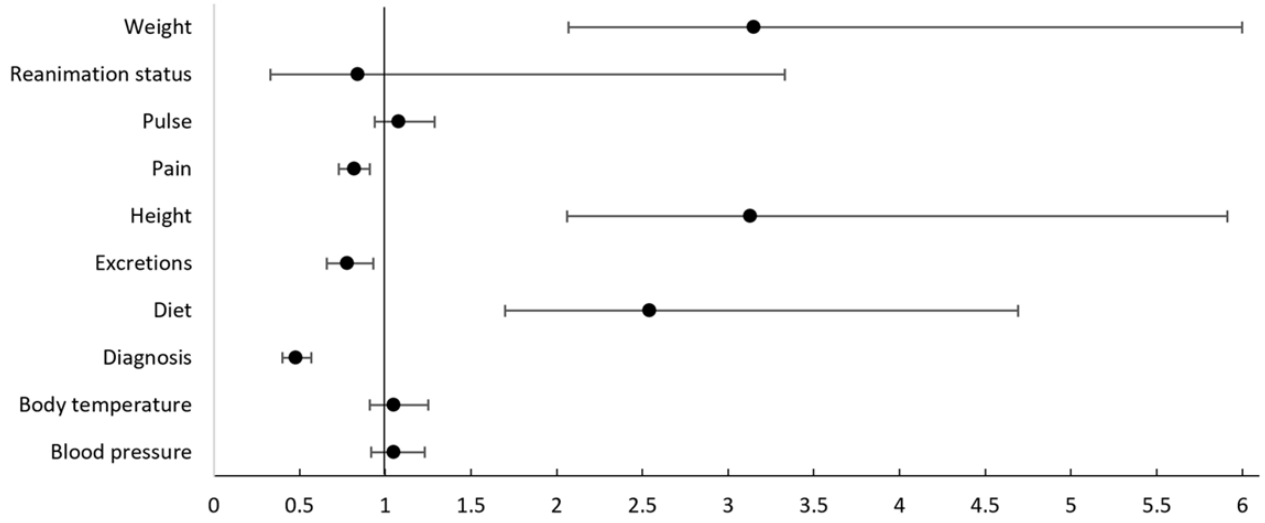
The mean number of documented items was 6.25 (SD 2.15) out of 10 in paper-based medical records and 7.13 (SD 2.01) out of 10 in EMRs. The Levene test confirmed the homogeneity of variances. The Shapiro-Wilk test did not confirm normal distributions, but the QQ plots show an approximation to a normal distribution and a comparable degree of normality (Multimedia Appendix 2). The unpaired *t* test confirmed the EMRs were statistically significantly more complete than the paper-based medical records under equal variances in the 2 record types (t_{178}=–2.469; *P*=.01; *d*=–0.428). When looking at the 10 items separately, data from chi-square tests showed that the documentation of diet increased from being present in 30% (13/44) of the paper-based medical record to 75% (102/136; *P*<.001) in the EMR, height from 27% (12/44) to 85.3% (116/136; *P*<.001), and weight from 27% (12/44) to 86% (117/136; *P*<.001). At the same time, documentation of diagnosis decreased from being present in 100% (44/44) of the paper-based medical records to 49% (66/136; *P*<.001) in the EMR, excretions from 86% (38/44) to 68% (92/136; *P*=.02), and pain from 95% (42/44) to 78% (106/136; *P*=.008). The documentation of vital signs such as blood pressure (*P*=.47), body temperature (*P*=.497), and pulse (*P*=.28) remained unchanged on a high level of completeness, while the documentation of reanimation status (*P*=.73) remained unchanged on a low level of completeness (Table 2). Positive relative risks (Figure 4) illustrate the association of the electronic record type (exposure) with complete documentation (outcome). The confidence intervals represent 5000 bootstrap replications.

Table 2. Change of completeness.

Variable	Type of record		Chi-square (*df*)	*P* value	RR[a] (95% CI)
	Paper (n=44), n (%)	Electronic (n=136), n (%)			
Blood pressure	37 (84)	120 (88.2)	0.5 (1)	.47	1.05 (0.92-1.23)
Body temperature	36 (81.8)	117 (86)	0.5 (1)	.497	1.05 (0.91-1.25)
Diagnosis	44 (100)	66 (48.5)	37.1 (1)	<.001	0.48 (0.40-0.57)
Diet	13 (29.6)	102 (75)	29.8 (1)	<.001	2.54 (1.70-4.69)
Excretions	38 (86.4)	92 (67.7)	5.8 (1)	.02	0.78 (0.66-0.93)
Height	12 (27.3)	116 (85.3)	54.5 (1)	<.001	3.13 (2.06-5.91)
Pain	42 (95.4)	106 (77.9)	7.0 (1)	.008	0.82 (0.73-0.91)
Pulse	36 (81.8)	120 (88.2)	1.2 (1)	.28	1.08 (0.94-1.29)
Reanimation status	5 (11.4)	13 (9.6)	0.1 (1)	.73	0.84 (0.33-3.33)
Weight	12 (27.3)	117 (86)	56.5 (1)	<.001	3.15 (2.07-6.00)

[a]RR: relative risk.

Figure 4. Forest plot of relative risks.



Discussion

Principal Findings and Comparison to Previous Work

The main findings of this study confirm an improved completeness of the analyzed information in the EMR on average. This provides further evidence for the suggestion that the general completeness of documentation can improve after the implementation of an EMR. The findings align with the results of similar studies, showing improvements in other data quality dimensions like the accuracy [26] or legibility [27] of documentation. However, when looking at the completeness of the analyzed 10 items in detail, the improvements can only be seen in 3 out of 10 items (diet, height, and weight), while 3 different items exhibited a deterioration in completeness (diagnosis, excretions, and pain). This links to the results of Coffey et al [28], who found 5 of their 11 analyzed items to be more complete while also proving 1 of their elements to be less complete. The reason for the variation in the change in completeness may lie in the mechanism of how information reaches the record. In the paper-based medical records, all information was documented by hand by the various professional groups. EMRs, on the other hand, offer technical features, for example, automatically obtaining information from other digital sources, like patients' health insurance data [29]. This was manifested as a possible mechanism by Jang et al [30], who showed improved completeness in the EMR for the automatically filled information but not for the manually documented ones.

The analysis shows that roughly every second EMR was missing the documentation of a diagnosis. This is a remarkable change, as it was present in every paper-based record (44/44, 100% vs 66/136, 48.5%). In the first place, it must be clarified that the diagnosis is determined by a physician who enters it into an independently run hospital information system (HIS). This documented diagnosis can also be a preliminary diagnosis, which is used for distribution to the clinical disciplines and is present for every admitted patient. The HIS was already in operation when medical staff was still using the paper-based preprints for documentation purposes. After the EMR's implementation, the HIS was still in operation along with the EMR. That being said, it is undisputed that during the paper-based period as well as the electronic period, a diagnosis was indeed present for the patients. In the paper-based period, the diagnosis was transferred manually from the HIS into the paper-based preprints, when a record for a recently admitted patient was prepared by a nurse. Since the HIS and the EMR are produced by different software developers, the diagnosis cannot be transferred automatically from the HIS into the EMR. Due to this noninteroperability of the 2 independent digital systems, the manual transfer is still necessary in the electronic period. With the drop of completeness in mind, this double documentation was accepted and carried out in the period of the paper-based record. In the electronic period, the described double documentation has decreased. One possible explanation is that the HIS was not automatically accessible, when an employee had the paper-based record at hand. With the introduction of the EMR, the availability of the EMR became synonymous with the availability of the HIS, since both are accessible from a computer. Therefore, the transfer of the diagnosis from the HIS to the EMR may no longer have been considered necessary. Nevertheless, the reason for this difference remaining unclear illustrates that the sole analyzation of completeness of the documentation alone does not provide sufficient information about the actual quality of the provided treatment. In that matter, it must also be highlighted that the record can contain additional qualitative data entries, like free texts, which might complement the analyzed quantitative information. This underlines that an insufficient quality of documentation does not necessarily allow conclusions to be drawn about the quality of care, and vice versa.

Brown [31] emphasizes this by cautioning people to always consider the circumstances under which people put information into the record before drawing conclusions. This is a major issue because the completeness of documentation might be biased due to aspects that do not directly derive from clinical care. On the one hand, the hospital's reimbursement for the delivered care depends on what is documented and might cause a possible strengthened thorough filling of certain fields [32]. On the other hand, the burden caused by documentation tasks is critically

heavy. It is responsible for a high prevalence of burnout among physicians and nurses [33]. Therefore, clinically or legally unnecessary documentation might be evaded [34]. However, even though complete documentation might neither necessarily arise from nor be essential for the delivery of excellent clinical care, it is likewise of concern under the aspect of big data analytics. In this regard, it would be desirable for the discussed diagnosis to indeed be present in the EMR, even if it already exists in the HIS. An automatic transfer of this information could help to prevent the burden on staff resulting from manual transmission and ensure a complete data set. This is an important point, as the insights gained from analyzing big data offer numerous opportunities, like data-based personalized care in diagnostics and therapy or the support of scientific activities, both with the chance of saving lives and reducing health care costs at the same time [7,35]. It is therefore indispensable to recognize the possibility of changes in documentation due to the implementation or adaptation of EMRs. Only with this attention will it become possible to optimize the documentation process with a focus on the various benefits for all stakeholders, like patients [6], practitioners [36], organizations [5], and society [7].

Strengths and Limitations

The German health care system, in which the study was conducted, was heavily strained by the high number of COVID-19 cases and the associated use of intensive care units during the study period. Especially the first measuring phase (November 2020) fell into the first pandemic year when many planned procedures were suspended to increase hospital capacities. For the first lockdown period in Germany (March 2020), a decrease in orthopedic surgeries is described by approximately 80% [37]. A lockdown-like situation was again declared during the first measuring phase [38], which probably explains the difference in treated patients over the 2 measuring phases (n_{Paper}=44 vs $n_{Electronic}$=136). However, the similarity between the coded ICD diagnoses over different years (Table 1) suggests that the proven changes in completeness of documentation are not due to significant changes in the studied patient sample, but a detailed sample description based on socioeconomical data is missing due to data protection regulations. On the other hand, there is a study assuming a positive influence of the pandemic on the completeness of documentation since an incomplete documentation might have led to repetitive contacts with the patient, which could have been avoided if the documentation would have been complete in the first place [39]. However, this cannot be verified in this paper due to the lack of further measuring phases. Within this given context, the generalizability of the presented results remains limited.

Further, limitations regarding the analyzed data set have to be stated. The chosen unpaired *t* test is theoretically based on the assumption of normal distributions. This could not be confirmed statistically for the mean completeness scores by the Shapiro-Wilk test. Although *t* test has been shown to be robust to a missing normal distribution [40] and the QQ plots (Multimedia Appendix 2) indicate an approximation to a normal distribution, the results could still be biased by the broken assumption.

Moreover, the analyzed data set is missing any information on which person was entering the documentation regarding which patient. On the one hand, it might be arguable that the same physicians or nurses were documenting during the first and also the second measuring phases. This circumstance would make the 2 compared measuring phases dependent samples, having an impact on the chosen statistical model. Since the analyzed data set is missing this information, the results might be biased regarding a possible dependent or independent sample. However, the time passed between the 2 measuring phases might have led to a change of the employees since the teaching status of the hospital results in many young physicians or nurses who do not necessarily stay on the same ward for a long time. Moreover, the hospital in which the study was conducted has a rotation system in which clinicians rotate hospital-wide across different wards of the same discipline. Those 2 facts let us assume that the 2 compared samples are indeed independent. However, the lack of information regarding the documenting individual is preventing the use of advanced tests like mixed effect models. These could equally consider the record type on the one hand and the possible documenting individuals on the other hand, potentially advancing the results' reliability. However, the 15-month interval from the implementation date of the EMR to the second data collection signifies that there is only little risk of any possible changes in documentation due to a bias from the described effects of preimplementational documentation training [41] since the employees indeed underwent software training before they were allowed to use the EMR. Therefore, the shown changes in completeness are, in fact, most likely due to the implementation of the EMR.

Conclusions

The results show that implementing EMRs can influence the completeness of documentation. A demonstrated improved completeness might also facilitate an improvement of the described outcomes that depend on documentation that is of high quality, like the availability [4] and analyzability of information [7,35], the coordination of care [5], or patient safety [6]. However, at the same time, the results show that a deterioration of completeness is also conceivable with the accompanied risks. This highlights the importance of understanding the underlying mechanisms that determine these changes. The knowledge may help stakeholders manage the implementation of new EMRs or the optimization of existing EMRs. Future research should address mechanisms that can improve documentation while simultaneously reducing the burden on practitioners caused by documentation tasks.

Acknowledgments

This work is funded by the German Federal Ministry of Education and Research (grant 01GP1906B). The sponsor had no influence on study design, data collection, analysis, or the writing process.

Data Availability

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

Authors' Contributions

FW and UK conceptualized the article. Data collection and analysis were performed by FW, supervised by UK. The original draft of the manuscript was written by FW, and all authors reviewed and edited previous versions of the manuscript and contributed to the interpretation of the data. All authors read and approved the final manuscript.

Conflicts of Interest

None Declared.

Multimedia Appendix 1
STROBE Checklist.
[PDF File (Adobe PDF File), 150 KB - medinform_v12i1e47761_app1.pdf]

Multimedia Appendix 2
Q-Q-Plots.
[PDF File (Adobe PDF File), 86 KB - medinform_v12i1e47761_app2.pdf]

References

1. Gopal G, Suter-Crazzolara C, Toldo L, Eberhardt W. Digital transformation in healthcare - architectures of present and future information technologies. Clin Chem Lab Med 2019;57(3):328-335 [FREE Full text] [doi: 10.1515/cclm-2018-0658] [Medline: 30530878]
2. Mangiapane M, Bender M. EMR Adoption Model (EMRAM). In: Mangiapane M, Bender M, editors. Patientenorientierte Digitalisierung im Krankenhaus. Wiesbaden: Springer Vieweg; 2020:33-39.
3. Jacob PD. Chapter 3 - Management of patient healthcare information: healthcare-related information flow, access, and availability. In: Gogia S, Novaes M, Basu A, Gogia K, Gogia S, editors. Fundamentals of Telemedicine and Telehealth. London: Academic Press; 2020:35-57.
4. Embi PJ, Weir C, Efthimiadis EN, Thielke SM, Hedeen AN, Hammond KW. Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators. J Am Med Inform Assoc 2013;20(4):718-726 [FREE Full text] [doi: 10.1136/amiajnl-2012-000946] [Medline: 23355462]
5. Vos JFJ, Boonstra A, Kooistra A, Seelen M, van Offenbeek M. The influence of electronic health record use on collaboration among medical specialties. BMC Health Serv Res 2020;20(1):676 [FREE Full text] [doi: 10.1186/s12913-020-05542-6] [Medline: 32698807]
6. Yanamadala S, Morrison D, Curtin C, McDonald K, Hernandez-Boussard T. Electronic health records and quality of care: an observational study modeling impact on mortality, readmissions, and complications. Medicine (Baltimore) 2016;95(19):e3332 [FREE Full text] [doi: 10.1097/MD.0000000000003332] [Medline: 27175631]
7. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014;2:3 [FREE Full text] [doi: 10.1186/2047-2501-2-3] [Medline: 25825667]
8. Dick RS, Steen EB, Detmer DE. The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition. Washington: National Academies Press; 1997.
9. 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;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
10. 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]
11. Wurster F, Fütterer G, Beckmann M, Dittmer K, Jaschke J, Köberlein-Neu J, et al. The analyzation of change in documentation due to the introduction of electronic patient records in hospitals: a systematic review. J Med Syst 2022;46(8):54 [FREE Full text] [doi: 10.1007/s10916-022-01840-0] [Medline: 35781136]

12. Montagna S, Croatti A, Ricci A, Agnoletti V, Albarello V, Gamberini E. Real-time tracking and documentation in trauma management. Health Informatics J 2020;26(1):328-341 [FREE Full text] [doi: 10.1177/1460458219825507] [Medline: 30726161]
13. Zargaran E, Spence R, Adolph L, Nicol A, Schuurman N, Navsaria P, et al. Association between real-time electronic injury surveillance applications and clinical documentation and data acquisition in a South African trauma center. JAMA Surg 2018;153(5):e180087 [FREE Full text] [doi: 10.1001/jamasurg.2018.0087] [Medline: 29541765]
14. Thoroddsen A, Ehnfors M, Ehrenberg A. Content and completeness of care plans after implementation of standardized nursing terminologies and computerized records. Comput Inform Nurs 2011;29(10):599-607 [FREE Full text] [doi: 10.1097/NCN.0b013e3182148c31] [Medline: 22041791]
15. McCamley J, Vivanti A, Edirippulige S. Dietetics in the digital age: The impact of an electronic medical record on a tertiary hospital dietetic department. Nutr Diet 2019;76(4):480-485 [FREE Full text] [doi: 10.1111/1747-0080.12552] [Medline: 31199071]
16. Karp EL, Freeman R, Simpson KN, Simpson AN. Changes in efficiency and quality of nursing electronic health record documentation after implementation of an admission patient history essential data set. Comput Inform Nurs 2019;37(5):260-265 [FREE Full text] [doi: 10.1097/CIN.0000000000000516] [Medline: 31094915]
17. Meier-Diedrich E, Davidge G, Hägglund M, Kharko A, Lyckblad C, McMillan B, et al. Changes in documentation due to patient access to electronic health records: protocol for a scoping review. JMIR Res Protoc 2023;12:e46722 [FREE Full text] [doi: 10.2196/46722] [Medline: 37639298]
18. Wiebe N, Varela LO, Niven DJ, Ronksley PE, Iragorri N, Quan H. Evaluation of interventions to improve inpatient hospital documentation within electronic health records: a systematic review. J Am Med Inform Assoc 2019;26(11):1389-1400 [FREE Full text] [doi: 10.1093/jamia/ocz081] [Medline: 31365092]
19. Emekli E, Coscun Ö, Budakoglu I, Kiyak YS. Clinical record keeping education needs in a medical school and the quality of clinical documentations. Konuralp Med J 2023;15(2):257-265 [FREE Full text] [doi: 10.18521/ktd.1259969]
20. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147(8):573-577 [FREE Full text] [doi: 10.7326/0003-4819-147-8-200710160-00010] [Medline: 17938396]
21. Beckmann M, Dittmer K, Jaschke J, Karbach U, Köberlein-Neu J, Nocon M, et al. Electronic patient record and its effects on social aspects of interprofessional collaboration and clinical workflows in hospitals (eCoCo): a mixed methods study protocol. BMC Health Serv Res 2021;21(1):377 [FREE Full text] [doi: 10.1186/s12913-021-06377-5] [Medline: 33892703]
22. Referenzdatenbank der Qualitätsberichte der Krankenhäuser. Gemeinsamer Bundesausschuss. 2023. URL: https://qb-referenzdatenbank.g-ba.de/ [accessed 2023-12-22]
23. Prior L. Repositioning documents in social research. Sociology 2008;42(5):821-836 [FREE Full text] [doi: 10.1177/0038038508094564]
24. Ranegger R, Hackl WO, Ammenwerth E. Implementation of the Austrian Nursing Minimum Data Set (NMDS-AT): a feasibility study. BMC Med Inform Decis Mak 2015;15:75 [FREE Full text] [doi: 10.1186/s12911-015-0198-7] [Medline: 26384111]
25. Toney-Butler TJ, Unison-Pace WJ. Nursing Admission Assessment and Examination. Treasure Island (FL): StatPearls Publishing; 2018.
26. Yadav S, Kazanji N, Narayan KC, Paudel S, Falatko J, Shoichet S, et al. Comparison of accuracy of physical examination findings in initial progress notes between paper charts and a newly implemented electronic health record. J Am Med Inform Assoc 2017;24(1):140-144 [FREE Full text] [doi: 10.1093/jamia/ocw067] [Medline: 27357831]
27. Muallem YA, Dogether MA, Househ M, Saddik B. Auditing the completeness and legibility of computerized radiological request forms. J Med Syst 2017;41(12):199 [FREE Full text] [doi: 10.1007/s10916-017-0826-0] [Medline: 29101478]
28. Coffey C, Wurster LA, Groner J, Hoffman J, Hendren V, Nuss K, et al. A comparison of paper documentation to electronic documentation for trauma resuscitations at a level i pediatric trauma center. J Emerg Nurs 2015;41(1):52-56 [FREE Full text] [doi: 10.1016/j.jen.2014.04.010] [Medline: 24996509]
29. Seroussi B, Bouaud J. The (Re)-Relaunching of the DMP, the French shared medical record: new features to improve uptake and use. In: Ugon A, Karlsson D, Klein GO, editors. Building Continents of Knowledge in Oceans of Data: The Future of Co-Created EHealth. Amsterdam: IOS Press; 2018:256-260.
30. Jang J, Yu SH, Kim CB, Moon Y, Kim S. The effects of an electronic medical record on the completeness of documentation in the anesthesia record. Int J Med Inform 2013;82(8):702-707 [FREE Full text] [doi: 10.1016/j.ijmedinf.2013.04.004] [Medline: 23731825]
31. Brown ML. Can't you just pull the data? The limitations of using of the electronic medical record for research. Paediatr Anaesth 2016;26(11):1034-1035 [FREE Full text] [doi: 10.1111/pan.12951] [Medline: 27747978]
32. Pruitt Z, Pracht E. Upcoding emergency admissions for non-life-threatening injuries to children. Am J Manag Care 2013;19(11):917-924 [FREE Full text] [Medline: 24511988]
33. Gesner E, Gazarian P, Dykes P. The burden and burnout in documenting patient care: an integrative literature review. Stud Health Technol Inform 2019;264:1194-1198. [doi: 10.3233/SHTI190415] [Medline: 31438114]

34. Saravi BM, Asgari Z, Siamian H, Farahabadi EB, Gorji AH, Motamed N, et al. Documentation of medical records in Hospitals of Mazandaran University of medical sciences in 2014: a quantitative study. Acta Inform Med 2016;24(3):202-206 [FREE Full text] [doi: 10.5455/aim.2016.24.202-206] [Medline: 27482136]
35. Batko K, Ślęzak A. The use of big data analytics in healthcare. J Big Data 2022;9(1):3 [FREE Full text] [doi: 10.1186/s40537-021-00553-4] [Medline: 35013701]
36. Ommaya AK, Cipriano PF, Hoyt DB, Horvath KA, Tang P, Paz HL, et al. Care-Centered clinical documentation in the digital environment: solutions to alleviate burnout. National Academy of Medicine. 2018. URL: https://nam.edu/care-centered-clinical-documentation-digital-environment-solutions-alleviate-burnout/ [accessed 2023-12-22]
37. Kapsner LA, Kampf MO, Seuchter SA, Gruendner J, Gulden C, Mate S, et al. Reduced rate of inpatient hospital admissions in 18 German University Hospitals during the COVID-19 lockdown. Front Public Health 2020;8:594117 [FREE Full text] [doi: 10.3389/fpubh.2020.594117] [Medline: 33520914]
38. Videokonferenz der Bundeskanzlerin mit den Regierungschefinnen und Regierungschefs der Länder am 28. Oktober 2020. Presse- und Informationsamt der Bundesregierung. 2020. URL: https://www.bundesregierung.de/resource/blob/997532/1805024/5353edede6c0125ebe5b5166504dfd79/2020-10-28-mpk-beschluss-corona-data.pdf?download=1 [accessed 2023-12-22]
39. Curtis CA, Nguyen MU, Rathnasekara GK, Manderson RJ, Chong MY, Malawaraarachchi JK, et al. Impact of electronic medical records and COVID-19 on adult goals-of-care document completion and revision in hospitalised general medicine patients. Intern Med J 2022;52(5):755-762 [FREE Full text] [doi: 10.1111/imj.15543] [Medline: 34580964]
40. Wilcox RR. Introduction to Robust Estimation and Hypothesis Testing. Amsterdam: Academic Press; 2011.
41. Prokosch HU, Ganslandt T. Perspectives for medical informatics. reusing the electronic medical record for clinical research. Methods Inf Med 2009;48(1):38-44. [Medline: 19151882]

Abbreviations

EMR: electronic medical record
HIS: hospital information system
ICD: International Classification of Diseases
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by J Hefner; submitted 31.03.23; peer-reviewed by S Veeranki, N Nekliudov, C Hudak; comments to author 23.07.23; revised version received 10.08.23; accepted 23.10.23; published 19.01.24.

Please cite as:
Wurster F, Beckmann M, Cecon-Stabel N, Dittmer K, Hansen TJ, Jaschke J, Köberlein-Neu J, Okumu MR, Rusniok C, Pfaff H, Karbach U
The Implementation of an Electronic Medical Record in a German Hospital and the Change in Completeness of Documentation: Longitudinal Document Analysis
JMIR Med Inform 2024;12:e47761
URL: https://medinform.jmir.org/2024/1/e47761
doi:10.2196/47761
PMID:38241076

Original Paper

Application of Failure Mode and Effects Analysis to Improve the Quality of the Front Page of Electronic Medical Records in China: Cross-Sectional Data Mapping Analysis

Siyi Zhan[1*], MMed; Liping Ding[1*], BSc; Hui Li[1], MMed; Aonan Su[1], MMed

Zhejiang Provincial People's Hospital, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Aonan Su, MMed
Zhejiang Provincial People's Hospital
No. 158, Shangtang Rd
Hangzhou, 310000
China
Phone: 86 18814885258
Email: suaonan_512917@126.com

Abstract

Background: The completeness and accuracy of the front pages of electronic medical records (EMRs) are crucial for evaluating hospital performance and for health insurance payments to inpatients. However, the quality of the first page of EMRs in China's medical system is not satisfactory, which can be partly attributed to deficiencies in the EMR system. Failure mode and effects analysis (FMEA) is a proactive risk management tool that can be used to investigate the potential failure modes in an EMR system and analyze the possible consequences.

Objective: The purpose of this study was to preemptively identify the potential failures of the EMR system in China and their causes and effects in order to prevent such failures from recurring. Further, we aimed to implement corresponding improvements to minimize system failure modes.

Methods: From January 1, 2020, to May 31, 2022, 10 experts, including clinicians, engineers, administrators, and medical record coders, in Zhejiang People's Hospital conducted FMEA to improve the quality of the front page of the EMR. The completeness and accuracy of the front page and the risk priority numbers were compared before and after the implementation of specific improvement measures.

Results: We identified 2 main processes and 6 subprocesses for improving the EMR system. We found that there were 13 potential failure modes, including data messaging errors, data completion errors, incomplete quality control, and coding errors. A questionnaire survey administered to random physicians and coders showed 7 major causes for these failure modes. Therefore, we established quality control rules for medical records and embedded them in the system. We also integrated the medical insurance system and the front page of the EMR on the same interface and established a set of intelligent front pages in the EMR management system. Further, we revamped the quality management systems such as communicating with physicians regularly and conducting special training seminars. The overall accuracy and integrity rate of the front page (P<.001) of the EMR increased significantly after implementation of the improvement measures, while the risk priority number decreased.

Conclusions: In this study, we were able to identify the potential failure modes in the front page of the EMR system by using the FMEA method and implement corresponding improvement measures in order to minimize recurring errors in the health care services in China.

(JMIR Med Inform 2024;12:e53002) doi:10.2196/53002

KEYWORDS

front page; EMR system; electronic medical record; failure mode and effects analysis; FMEA; measures

Introduction

The electronic medical record (EMR) system is the main carrier of medical information that has details about the whole process of a physician's treatment for a patient [1]. The information on the front page of the EMR is condensed, which includes a patient's basic information, disease diagnosis, information on surgical or invasive operations, and medical expenses [2]. Since January 1, 2013, almost all tertiary hospitals in China have submitted the front pages of the EMRs of inpatients to the Hospital Quality Monitoring System led by the Bureau of Medical Administration and Medical Service Supervision and National Health and Family Planning Commission of the People's Republic of China [3]. The quality and management of the front pages of EMRs are critical for their application in medical services [4], research [2,5], education [6], and hospital management [7]. For example, some indicators for assessing the capacity of hospital medical services, such as the services for surgery and disease diagnosis, often utilize the information through the front page of the EMR for statistical purposes. However, there are many difficulties in the management of the front page of EMR. A survey conducted by the National Medical Record Management Quality Control Center of China [8] showed that more than 230 million front pages of EMRs in 2020 are established in China. Each of them contain over 100 fields. However, there are only 2.5 full-time coders on average in each hospital among 5439 medical institutions, and only 67.9% of them perform special quality control, while 24.2% of them use information technology to control the quality of the front page of the EMR system.

For reforming the medical insurance payment methods in China, the Chinese State Council's version of health insurance issued a notice in 2019 on the issuance of technical specifications and grouping schemes for the national pilot of diagnosis-related grouping payments for diseases [9]. Therefore, the front page of an EMR needs to be uploaded on the websites of the Health and Wellness Committee and the Health Insurance Authority, which means coders need to edit a front page twice to meet the different needs of both the sectors. The former is for hospital performance evaluation and the latter is for patient health insurance payment. The introduction of this policy in 2019 increased the difficulty of medical record management.

Failure mode and effects analysis (FMEA) is a proactive risk management tool that originated in the US military in the 1940s. It is widely applicable to human, equipment, and system failure modes, as well as hardware and software programs. FMEA finds out all the potential failure modes in a system and analyzes their possible consequences by mapping the subsystems and each subprocess that makes up the process one by one in the product design stage and process design stage [10]. Thus, the advantage of FMEA is that problems can be identified and improved during the system development phase to avoid possible problems. Moreover, the costs incurred to address software defects and failures at an early stage are lower compared to those incurred to address defects at a later stage. Initially, FMEA was widely used in engineering [11], food safety management [12], financial management [13], and so on. Thereafter, with the rising demands in health care services, FMEA was used for proactive health care risk analysis. Doctors often use the EMR system to record patients' visits. Any issue in the EMR system can affect the patient's visit process and visit records. According to a systematic review [14], 158 studies published from 1998 to 2018 and classified under 4 categories, namely, health care process, hospital management, hospital informatization, and medical equipment and production, reported the use of FMEA in health care systems for proactive health care risk evaluation. In FMEA, the risk priority number (RPN) is calculated by giving a numerical value (scoring) for the severity, frequency, and detectability of the risks or failures, which enable risk assessment of the system [10]. An EMR system named Heren (Zhejiang Heren Technology Corporation), which is installed in many hospitals in China, is used by physicians and medical record management coders and quality controllers for filling out the front page. The purpose of this study was to identify the possible failures in the front page data of the EMR and their causes and effects and to propose specific improvement measures to minimize errors. Moreover, we aimed to compare the EMRs before and after introducing the measures to verify the efficacy of the improvement measures. For this, we reviewed previous relevant literature through PubMed, Embase, Web of Science, and Cochrane Library. During this review, we found that although FMEA has been used in some studies for improvement of some facets of EMRs, no study has used FMEA for improving the efficiency the front page of the EMR [15,16]. Thus, to the best of our knowledge, ours is the first study to apply FMEA to identify the potential failures on the front page of the EMR in China and the causes and effects of these failures and to perform a before-and-after comparison of the revised front page of the EMR.

Methods

Study Design

We conducted a cross-sectional study from January 1, 2020, to May 31, 2022, in Zhejiang People's Hospital, which is one of the largest public hospitals in Zhejiang province with more than 100,000 hospital discharges per year. During the period of our research, the number of hospital discharges reached 250,774, which means the same number of front pages of EMRs needed to be filled and coded.

Steps of FMEA

Assembling a Panel for FMEA

Ten experts, including clinicians, medical record coders, and hospital administrators, were invited to assess the potential risks of the EMR system in China. Since coders and quality controllers were necessary to ensure the accuracy of the front page of the EMR, only those who had been working full-time on this task for more than 5 years and who had achieved a coding accuracy rate of more than 95% and who had checked more than thousands of medical records for quality were included. Before we began our study, the organizer introduced the theme of our study to ensure that every expert knew the process of FMEA and the importance of a front page of an EMR. Then, the time and place for each discussion was planned to ensure that the process ran smoothly.

Mapping the Process and Subprocesses

Each expert mapped the process and subprocess of completing a front page of an EMR alone initially to avoid interference from others. For example, there are 2 data sources for the content on the front page of the EMR: information automatically imported from the hospital information system that is mainly used by physicians and information that is filled in manually by the physician. Thus, different experts could map their own process according to their work experience. Thereafter, all experts were gathered to draw the final process and subprocess to achieve the completeness of the whole system.

Brainstorming to Identify Potential Failure Modes in Each Subprocess and Their Causes and Effects

The implementation process of this step is consistent with the mapping process. At first, each expert could think about every potential failure mode individually. Then, all the experts summarized all the modes and discussed many more potential failure modes by brainstorming once again. In addition, the views on effects and reasons for failure modes were exchanged by experts. Since there were so many issues that could result in potential failure modes, our team summarized the main causes and created a questionnaire for randomly selected physicians and coders to answer.

Calculating the RPN

A scoring criterion was used to evaluate the severity, frequency, and detectability of the failures, and each dimension was divided into 10 points. Then, the RPN was calculated by using the score of the 3 dimensions (RPN = Severity × Frequency × Detectability) to evaluate the final score of each failure mode, which ranges from 0 to 1000. To improve the consistency and accuracy of scoring, the rating weight of each expert was based on their professional title grade, work experience, and familiarity with FMEA. In addition, a risk assessment criterion was established to avoid any dispute about the scores given by the experts.

Proposing Improvement Measures for Each Failure Mode

Since a low RPN could result from severity, frequency, or detectability and a low score for each dimension could be caused by many different reasons, it is necessary to find out the main issues. According to the Pareto principle, 80% of the consequences are due to only 20% of the potential causes [17]. Our team used the Pareto principle to identify the pressing causes that need to be addressed. Then, the experts proposed one or more corresponding improvement measures for each failure mode. Further, the feasibility and effectiveness of improvement measures were also discussed.

Comparing the Quality Before and After the Improvements

The experts evaluated the quality of the front page of the EMR before and after the application of the improvement measures. The RPN score was bound to improve if these improvement measures were effective.

Ethical Considerations

This study did not involve any patient data or ethical data, and the ethics approval committee of Zhejiang Provincial People's Hospital specified that no ethics approval was required.

Statistical Analysis

We performed statistical analyses using SPSS (version 20.0; IBM Corp). Two-sided *t* tests were performed to compare the RPNs of the front page of the EMR before and after applying the improvement measures. *P* values <.05 were considered statistically significant.

Results

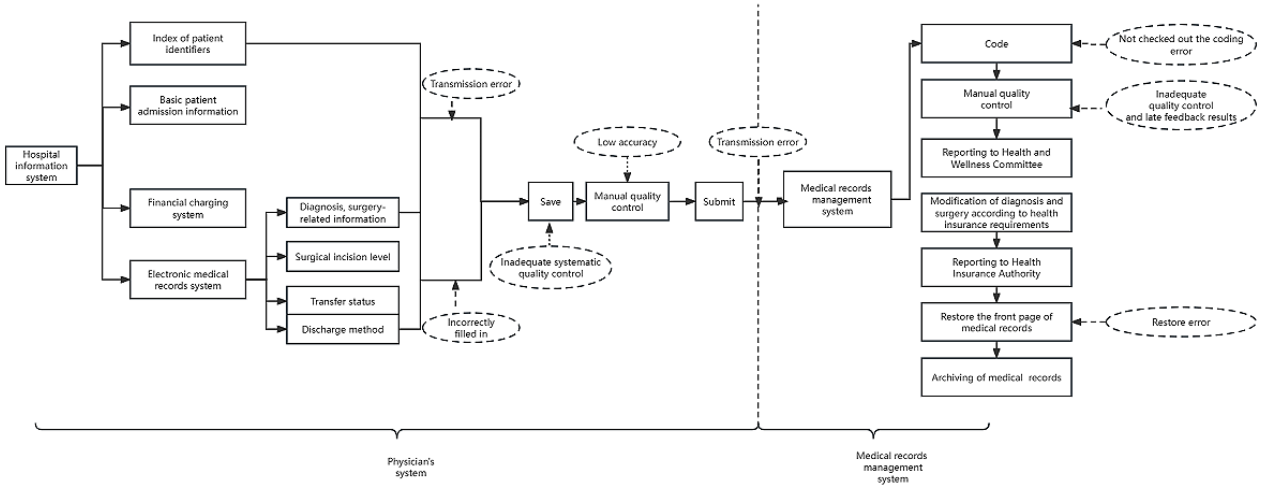
Assembling a Panel for FMEA

Our expert panel consisted of 10 experts in 5 different fields. There were 4 physicians, 2 coders, 2 hospital administrators, 1 quality control staff, and 1 information engineer who expressed different views and opinions on the front page of the EMR.

Mapping the Process and Subprocesses of the Front Page of EMR

The expert panel identified 2 main process steps and 6 subprocesses after discussion (Figure 1). The 2 main process steps were management of the physician’s system and management of the medical record system. The 6 subprocesses were information import from physicians’ hospital information system, front page filled by physicians, front page quality control by physicians, transmission of information in the EMR system, coders’ proofreading and coding, and front page quality control by EMR management.

Figure 1. Process map of the front page of the electronic medical record system.



Brainstorming to Identify Potential Failure Modes in Each Subprocess and Their Causes and Effects

The front page of the EMR that was evaluated in this study is shown in Figure 2. According to the process map of the front page of the EMR, the expert panel found that there were 13 potential failure modes, which can be mainly divided into 2 categories. One category is the low accuracy in a variety of information, including basic patient information, treatment information, and cost information. The other category is the low detection in a variety of information, including incorrect case header coding, incomplete quality control, and transmission errors. Regarding the causes for the failures, 115 physicians and coders filled out a questionnaire and summarized 15 main causes. The main causes were incompleteness of the front page, error in information, or incorrect diagnosis-related group, which is risky for hospital medical quality management, academic research, and medical insurance payment (Table 1).

Figure 2. Front page of the electronic medical record.

The Front Page of the Medicla Record of Inpatients

Medical Payment Options:			Medical institution code:			
Health Card Number:	No. of Hospitalizations:		Medical Card Number:			
Name:	Sex:	Birthday:	Age:	Nationality:		
Age(Less Than One Year Old):	Newborn Birth Weight:	Newborn Admission Weight:				
Birthplace:	Registered Birthplace:	Ethnological:				
I.D. Number:	Occupation:	Matrimonial:				
Current Address:	Telephones:	Postcode:				
Residential Address:		Postcode:				
Workplace And Address:	Unit Phone Number:	Postcode:				
Contact Name:	Relationship:	Contact Person'S Address:	Telephones:			
Pathway to Hospitalization:						
Date of Admission:	Admission Department:	Admission Ward:	Referral Unit:			
Date of Discharge:	Discharge Department:	Discharge Ward:	Days of Hospitalization:			
Outpatient Diagnosis:			Disease Codes:			
Discharge Diagnosis:	Disease Codes::	Admission Condition:	Discharge Diagnosis:	Disease Codes:	Admission Condition:	
Admission Condition: 1. Yes; 2. Clinically Undetermined; 3. Unknown; 4. None						
External Causes of Injuries, Poisoning:			Disease Codes:			
Pathological Examinations:			Pathological Number:			
Drug Allergy:						
Blood Type:	Rh:					
Chairman of Section:	Chief (Deputy) Physician:	Attending Doctor:	Resident Doctor:			
Student Nurse:	Refresher Doctor:	Intern:	Coder:			
Quality of Medical Record:	Quality Control Doctor:	Quality Control Nurse:	Quality Control Date:			
Surgery:						
Surgical Coding:	Date of Surgery:	Surgical Level:	Name of Surgery: Operator, First Assistant, Second	Incision Healing Grade:	Anesthesia:	Anesthesiologist
Methods of Discharge:	Whether there Is A Plan to be Admitted to the Hospital After 31 Days of Discharge			Coma Time In Patients With Craniocerebral Injuries:		
Hospitalization Expenses (Yuan):	Total Cost (Yuan):	(Out-of-Pocket Expenses) :				
1. Comprehensive Medical Service	(1) General Medical Service Expenses:	(2) General Treatment Operation Expenses:		(3) Nursing Care Expenses:	(4) Other Expenses:	
2. Diagnostic Category:	(5) Diagnostic Pathology Expenses:	(6) Laboratory Diagnostic Expenses:		(7) Diagnostic Imaging Expenses:	(8) Clinical Diagnostic Program Expenses:	
3. Therapeutic Category:	(9) Non-Surgical Item Treatment Expenses:	(Clinical Physiotherapy Expenses:)				
	(10) Surgical Treatment Expenses:	(Anesthesia Expenses; Surgical Expenses:)				
4. Rehabilitation:	(11) Rehabilitation Expenses:					
5. Traditional Chinese Medicine:	(12) Chinese Medicine Treatment Expenses:					
6. Western Medicines:	(13) Western Medicine Expenses	(Antimicrobial Drug Expenses:)				
7. Traditional Chinese Herb:	(14)Chinese Patent Medicines Expenses:	(15)Chinese Medicinal Herb:				
8. Blood Products:	(16) Blood Expenses:	(17) Albumin-Based Products Expenses:		(18) Expenses for Globulin-Based Products:	(19) Coagulation Factor-Based Products Expenses:	(20) Cytokine-Based Products Expenses:
9. Consumables :	(21) Charges for Disposable Medical Materials for Examinations :	(22) Charges for Disposable Medical Materials for Therapeutic Use :		(23) Charges for Surgical Disposable Medical Materials		
10. Other Categories :	(24) Other Expenses:					
Diagnostic Compliance:	Outpatient & Discharge:	Admission & Discharge:	Pre-Operative And Post-Operative:	Radiation And Pathology:	Clinical And	
0. Not Done; 1. Conform; 2. Not Conform; 3. Not Sure						
Single-Case Management:	Clinical Pathway Management:	Resuscitation:	Outcome Situation:			
Description:						
(A) Medical Payment Methods 1. Basic Medical Insurance for Urban Workers 2. Basic Medical Insurance for Urban Residents 3. New Rural Cooperative Medical Care 4. Poverty Relief 5. Commercial Medical Insurance 6. Full Public Expense 7. Full Self-Funding 8. Other						
(B) Where the Hospital Information System Can Provide A List of Inpatient Expenses, the Front Page of the Inpatient Medical Record May Not be Filled In With "Inpatient Expenses"						

Table 1. Potential failure modes with their causes and effects.

Process	Failure modes	Reasons	Effects
Transmission in the hospital information system	• Basic information transmission error • Inpatient information transmission error • Expenses information transmission error	• Data interface errors	• The original data on the front page are erroneous • The DRG[a] is erroneous • Affects patients' medical reimbursement
Front page filled by physicians	• Incorrectly filled-in medical information • Incorrectly filled in other information	• Do not understand the filling criteria • Do not fill in carefully • Incomplete quality control reminders	• The original data on the front page are erroneous • The DRG is erroneous • Affects patients' medical reimbursement
Front page quality control by physicians	• Inadequate quality control • Inaccurate quality control	• No emphasis on quality control • Unfamiliar with quality control rules • Complexity of quality control rules • Lack of information assistance	• The original data on the front page are erroneous • The DRG is wrong • Affects patients' medical reimbursement
Transmission in the physicians' EMR[b] system	• Inconsistency between the received data in the EMR system and original data	• Data interface errors • Encoding conversion error	• The original data on the front page are wrong • The DRG is wrong
Coders' proofreading and coding	• No data errors were found • Wrong code for diagnosis, surgery, or operation • Restoration error • Diagnostic and surgical operation codes do not meet the requirements of patients' insurance	• Formal quality control rules are too simple • Lack of internal quality control reminders • Insufficient professional capacity of coders • Few training opportunities for coders • Inadequate communication between coders and doctors • The criteria are different between the requirements of patients' insurance and front page	• Erroneous data persist • The DRG is wrong
Front page quality control by EMR management	• Inadequate quality control • Late feedback for the results of quality control	• Using a sampling model to conduct quality control • Insufficient professional capacity of quality control staff • Complexity of quality control rules • Lack of information assistance	• Unable to find all errors on the front page • Erroneous data persist • The DRG is wrong

[a]DRG: diagnosis-related group.

[b]EMR: electronic medical record.

Calculating the RPN

Before calculating the RPN, a risk assessment criterion was established to evaluate the quality of the front page of the EMR (Table 2).

The rating weight of each expert was calculated to reduce the influence caused by the individual subjective factors of the experts (Table 3).

Table 2. Risk assessment criteria for the quality management of the front page of the electronic medical record system.

Grade	Severity	Criteria for risk severity	Frequency	Criteria for risk frequency	Detectability	Criteria for risk detectability
10	Very high	Make the score of the front page of the EMR[a] below 20	Extremely high	Every time	Very low	Cannot be detected
9	Very high	Make the score of the front page of the EMR between 20 and 30	Very high	Almost every time	Very low	Hard to detect
8	High	Make the score of the front page of the EMR between 30 and 40	Very high	One time every half day	Low	Seldom detected
7	High	Make the score of the front page of the EMR between 40 and 50	High	More than one time every day	Low	Seldom detected
6	Middle	Make the score of the front page of the EMR between 50 and 60	High	More than one time every week	Middle	Easy to be detected
5	Middle	Make the score of the front page of the EMR between 60 and 70	Middle	More than one time every month	Middle	Easy to be detected
4	Middle	Make the score of the front page of the EMR between 70 and 80	Middle	More than one time every year	High	Very easy to be detected
3	Low	Make the score of the front page of the EMR between 80 and 90	Low	One time every year	High	Very easy to be detected
2	Low	Make the score of the front page of the EMR between 90 and 100	Very low	Less than one time every year	High	Very easy to be detected
1	Very low	Does not affect the score of the front page of the EMR	Extremely low	Never	Very high	No failure modes

[a]EMR: electronic medical record.

Table 3. Details of the expert panel.

Position	Rating weight	Working experience (years)	Familiarity with FMEA[a]	Rating weight
Physician	High	>20	General	9/10
Physician	Middle	10-20	General	7/10
Physician	Middle	1-5	Familiar	6/10
Physician	Primary	6-10	General	5/10
Coder	High	10-20	Familiar	9/10
Coder	Middle	6-10	Familiar	7/10
Quality control staff	High	10-20	Not very familiar	7/10
Administrator	High	>20	Not very familiar	8/10
Administrator	Middle	10-20	Familiar	8/10
Information engineer	Primary	1-5	Familiar	5/10

[a]FMEA: failure mode and effects analysis.

Proposing Improvement Measures for Each Failure Mode

According to the principle of Pareto, there were 7 causes in our study that contributed to 80% of the consequences (Figure 3), which can be addressed by revamping the information and quality management. For example, we integrated the medical insurance system with the front page of the EMR on the same interface and established a set of intelligent front pages for the EMR management system. In addition, we revamped the management of quality, such as communicating with physicians regularly and conducting special training seminars (Table 4).

Figure 3. The 7 causes that contributed to 80% of the failure modes in the electronic medical record system, according to the principle of Pareto. QC: quality control.

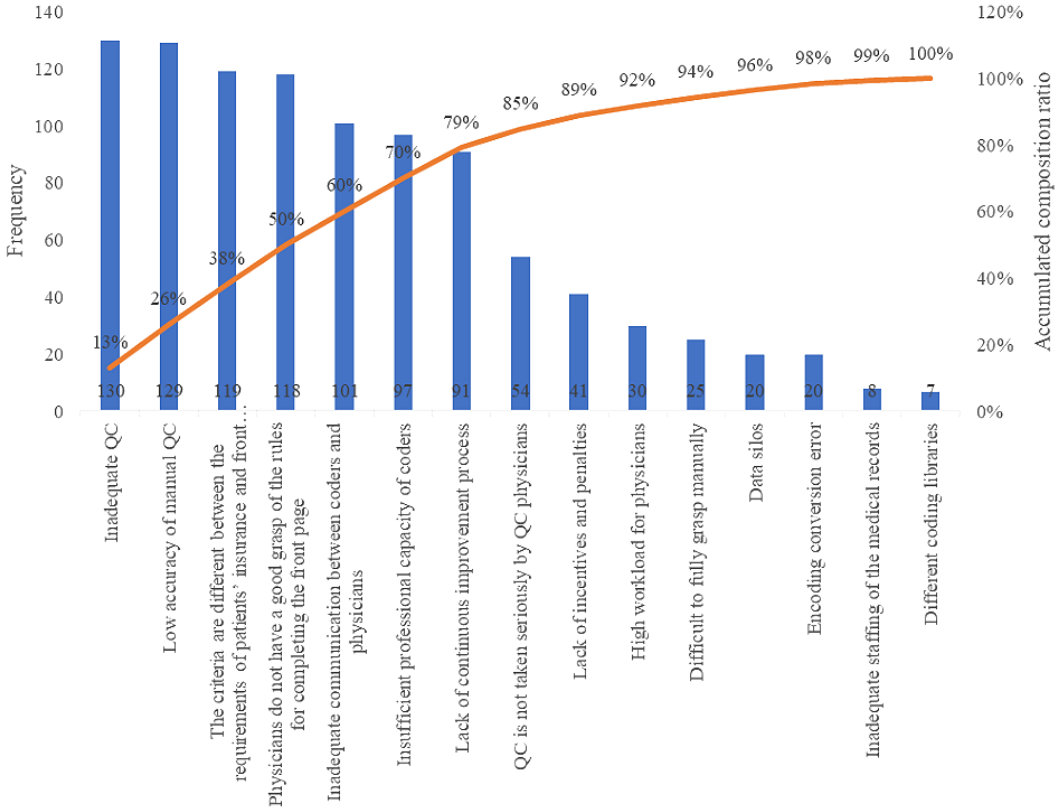


Table 4. Improvement measures for the causes of potential failure modes.

	Why	What	How and When	Where	Who
Revamp the information	• The rules for quality control are inadequate • Low accuracy in manual quality control • The criteria are different between the requirements of patients' insurance and front page details	• Establish a set of intelligent front pages in the EMR[a] management system	• Added quality control rules from May to September 2021 • Embedded quality control systems into the physician's system and front page system from July to October 2021 • Integrated medical insurance system and the front page of the EMR from July to December 2021	• EMR center and information center	• Staffs of EMR center and information engineers
Revamp quality management	• Formal quality control rules are simple • Insufficient professional capacity of coders • Inadequate communication between coders and physician • Lack of a continuous improvement process	• Improve the correctness of the front page by physician • Improve the professional capacity of coders • Establish effective communication channels • Form a continuous improvement process	From January to May 2021 • Conducted special training seminars • Prepared a quality and information management manual • Implemented professional training regularly • Invited experts for guidance • Communicated with physicians regularly • Provided timely feedback to physicians on quality control results	• EMR center and clinical departments	• Staffs of EMR center, information engineers, and physicians

[a]EMR: electronic medical record.

Comparing the Front Page of EMR Before and After Improvement

Before carrying out improvement measures, the highest RPN given by the experts was 296.3 for the failure mode "wrong code for diagnosis, surgery, or operation," which was due to the quality rules being too simple, while the lowest was 50.6 for "basic information transmission error," which was caused by wrong data interface or conversion error. On average, the final RPN was 181.2. The highest score for severity was for "wrong code for diagnosis, surgery, or operation" and the lowest was for "expense information transmission error." The highest score for frequency was for "incorrectly filled-in medical information" and the lowest was for "expense information transmission error." As for detectability, the highest score was for "wrong diagnosis-related group code of diagnosis, surgery, or operation," "restoration error," and "late feedback for quality control results." The lowest score for detectability was for "basic information transmission error." Our team calculated the RPN of the revised front page of the EMR after implementing the improvement measures mentioned in Table 5, and the final RPN was 95.0, which was lower than that of the original front page of the EMR (RPN=181.2).

The RPN of every failure mode decreased after implementing the improvements, and the mode for the late feedback for quality control decreased the most remarkably (Table 5). In addition, the accuracy rate of the basic information (χ^2_1=269.6; P<.001); inpatient information (χ^2_1=175.9; P<.001); diagnosis, surgery, and operation code (χ^2_1=32.9; P<.001); and the overall accuracy rate of the front page (χ^2_1=239.3; P<.001) and the integrity rate of the front page (χ^2_1=110.4; P<.001) increased significantly (Table 6).

Table 5. Comparison of the risk analysis before and after failure mode and effects analysis model improvement.

Process, failure modes	Before FMEA[a]				After FMEA			
	Severity	Frequency	Detectability	Risk priority number	Severity	Frequency	Detectability	Risk priority number
Information import from HIS[b]								
Basic information transmission error	3.1	3.4	4.8	50.6	3.1	3.1	3.7	35.6
Inpatient information transmission error	3.9	2.3	6.6	59.2	3.6	2.1	6.2	46.9
Expenses information transmission error	2.4	2.1	6.6	33.3	2.2	2.0	6.0	26.4
Front page filled by physicians								
Incorrectly filled-in medical information	6.1	6.6	6.4	257.7	5.8	6.2	6.0	215.8
Incorrectly filled-in other information	2.7	5.8	6.1	95.5	2.3	5.1	5.5	64.5
Front page quality control by physicians								
Inadequate quality control	6.5	6.4	6.4	266.2	6.0	5.1	5.2	159.1
Inaccurate quality control	6.1	6.5	6.5	257.7	4.9	5.0	4.7	115.2
Information import from the physician's EMR[c] system								
Inconsistency between the received data in EMR system and the original data	3.6	3.3	5.3	63.0	3.1	3.1	4.7	45.2
Coders' proofreading and coding								
No data errors were found	5.2	6.3	6.3	206.4	4.5	5.3	4.8	114.5
Wrong code for diagnosis, surgery, or operation	6.7	6.6	6.7	296.3	4.6	4.5	5.0	103.5
Restoration error	6.6	6.1	6.7	269.7	4.4	4.6	4.7	95.1
Front page quality control by EMR management								
Inadequate quality control	6.1	6.5	6.0	237.9	4.6	5.1	4.6	107.9
Late feedback of quality control results	6.1	6.4	6.7	261.6	4.7	4.9	4.6	105.9
Average	N/A[d]	N/A	N/A	181.2	N/A	N/A	N/A	95.0

[a]FMEA: failure mode and effects analysis.

[b]HIS: hospital information system.

[c]EMR: electronic medical record.

[d]N/A: not applicable.

Table 6. Comparison of the accuracy and integrity of the front page of the electronic medical records before and after failure mode and effects analysis model improvement.

Items	Front pages (n)	Accuracy rate of basic information	Accuracy rate of inpatient information	Accuracy rate of diagnosis, surgery, and operation code	Overall accuracy rate of front page	Integrity rate of front page
Before	48,509	94.09	95.28	97.29	93.44	96.15
After	78,890	96.09	96.74	97.81	95.48	97.26
Chi-square *(df)*	N/A[a]	269.6 (1)	175.9 (1)	32.9 (1)	239.3 (1)	110.4 (1)
P value	N/A	<.001	<.001	<.001	<.001	<.001

[a]N/A: not applicable.

Discussion

The quality of the front page of an EMR is quite important not only for hospital performance management [2] but also for insurance payments to patients [15]. Thus, it is necessary to improve the effectiveness of the front page of the EMR. There are many risk management tools for investigating the potential problems in an EMR system, such as Expert Delphi [18], scenario analysis method [19], and SWOT (strengths, weaknesses, opportunities, and threats) analysis method [20]. The advantage of Expert Delphi is that everyone's opinions are collected and that of scenario analysis is that it identifies risks by designing multiple possible future scenarios. The advantage of SWOT is that it identifies the strengths, weaknesses, opportunities, and costs of the project, thus qualitatively identifying the project risks from multiple perspectives. FMEA is a risk management tool that has most of the advantages of the above tools. FMEA can not only change the occurrence of risk from postprocessing to preemptive prevention but is also a simple and a practical risk quantification method [10]. In recent years, FMEA has been widely used in various fields, including the medical field. Studies on medical services [21], medicine distribution [22], infection control [23], and medical equipment operation and maintenance [24] have used FMEA to date.

In this study, we found potential failures existing in the EMR system of China and proposed improvement measures to solve the problems by using FMEA. Our results showed that there were 2 main processes and 6 subprocesses in the EMR system that showed 13 potential failure modes. The 2 main process steps were management of the physician's system and management of the medical record system. The 6 subprocesses were information transmission in the hospital information system, front page filled by physicians, front page quality control by physicians, information transmission in the EMR system, coders' proofreading and coding, and front page quality control by EMR management. This finding is similar to that reported in a study performed in Indonesia [25], wherein potential failure modes included incomplete or missing medical record files, mistakes caused by coders, and excessive code writing or upcoding [25].

According to the principle of Pareto and from questionnaire responses, we found that there were 7 causes in our study that contributed to 80% of the consequences, which can be divided into 2 aspects for the resolution of errors. One aspect was to revamp the information by establishing a set of intelligent front pages in the EMR management system to solve the problems of inadequate information and inaccurate quality control and to implement different codes of management or payment. In this study, we established quality control rules for medical records and embedded them in the system first. Accurate quality control rules are important for maintaining data quality. For example, Carlson et al [26] used quality control rules to identify common logical problems, including incomplete data, invalid values, and inconsistent data, in a clinical data set of an intensive care unit. Hart et al [27] reported >50% decrease in rejected records across patient information, service information, and financial information in 6 months by using quality rules. In addition, we integrated the medical insurance system with the front page of the EMR on the same interface. The other aspect was to revamp the management of quality by conducting multichannel trainings for doctors and coders, creating a quality and information management manual, and communicating with physicians and coders regularly. Previous studies [28-30] have shown a high rate of errors in physician coding for professional services, which can be risky in medical care services. One study [31] showed that clinicians and coders differ in their understanding of disease coding and need to communicate in a timely manner. Some of our measures are also consistent with those previously reported [25] that a hospital needs to update coding training for coders and provide guidance and validation of coding for physicians as well.

After implementing improvement measures, we found that the RPN of every failure mode decreased. The most significant decline in RPN was for the mode on the late feedback for quality control results. Many studies have proved the benefits of artificial intelligence. For example, machine learning could improve the content of medical records by identifying patients' medical information [32] or by predicting the onset of disease [33]. Therefore, we applied artificial intelligence to establish an intelligent front page for the EMR management system and then embedded it in the doctor's medical record writing interface and medical record quality control interface, which made it possible for real-time quality control of the front page. The second indicator of decline was inaccurate quality control of the front page by physicians. The original data on the front page, such as basic patient information, expenses, and surgery, are filled by physicians who decide the quality of the front page mostly [34]. After the amendments, the accuracy and integrity

of the front page were both improved for those measures, which helped the diagnosis-related group to be more specific and the evaluation of the hospital performance more precise. In addition, the quality of the front page of EMR is quite important for patients. A complete front page of the medical record enables doctors to grasp important information about the patient in a short period, such as family history, allergy history, and important test results and facilitates doctors to quickly and accurately judge the patient's condition and formulate diagnosis and treatment plans, thereby reducing overmedication and even erroneous medical treatment.

Human factors engineering and user-centered designs are indispensable components of mobile health technology design and implementation [35]. Human factors emphasize human needs and capabilities as the core of the design technology system, making people the most important consideration in the design process and aiming to achieve the goal of "making machines fit people." Regarding EMR system update, physicians, medical record coders, and quality controllers are the target users, and they will resist the technology when they believe it does not meet their expectations and needs [36]. For this reason, this study was conducted through brainstorming and questionnaires to inform the needs of physicians, coders, and others regarding the front page of the EMR system. For example, incorrectly filled-in medical information and quality control proposed by physicians, coders, and other users prompted engineers to establish a set of intelligent front pages in the EMR management system. The usability of the EMR system is evaluated by its effectiveness, efficiency, and suitability for target users. Although we did not use questionnaires to analyze the satisfaction of doctors, coders, and others with the improved EMR system, the results of FMEA showed that RPN was greatly reduced after the system was improved; thus, it can be hypothesized that the user's satisfaction with the system has been enhanced. Moreover, the overall accuracy rate of the front page (P<.001) and the integrity rate of the front page (P<.001) were significantly enhanced after implementing the improvement measures, thereby demonstrating the increase in the effectiveness of the system. The number of front pages of EMR increased from 48,509 to 78,890 with the same amount of time and labor, which proves that the efficiency of the system was also improved.

Our study has several strengths. First, medical research FMEA has mostly been performed for health care processes, hospital management, etc. For example, a study performed in Sri Lanka used FMEA to improve medication safety in the dispensing process [22], while another study aimed to increase the efficiency and success rate of patients with acute ischemic stroke [25]. No study has used FMEA for improving the front page of EMR in China before. Therefore, this is the first study performed in China, which can provide the base for future studies. Second, most studies only used FMEA to find potential failure modes and propose improvement measures, but the system was not evaluated after the application of those measures. However, our study used FMEA to compare the RPN of the front page of the EMR before and after applying the improvement measures to verify the efficiency of the system. Our study also has some limitations. The first limitation was that the method we used is not advanced since there are many better methods such as data envelopment analysis [37] and fuzzy RPN method [38]. The second limitation was that the process of scoring the system by the experts was subjective although we had set weights for the experts' scores. The third limitation was that we did not use additional methods to validate the results, which we aim to improve in the future. Lastly, although the EMR system called Heren has been used in many hospitals, different hospitals may use different types of Heren. Consequently, the generalizability of this study and the findings should be considered cautiously. In conclusion, we improved the front pages of the EMRs in China based on the potential failure modes found by the FMEA method.

Acknowledgments

This project was supported by Project of 2023 Zhejiang Province Archives Science and Technology Project Research (2023-34).

Data Availability

Data sets are available from the corresponding author on reasonable request.

Authors' Contributions

SZ wrote the main manuscript. AS, HL, and SZ prepared the figures and tables. AS, HL, SZ, and LD designed the study. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20;Suppl 1:S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
2. Wu C, Zhang D, Bai X, Zhou T, Wang Y, Lin Z, China PEACE Collaborative Group. Are medical record front page data suitable for risk adjustment in hospital performance measurement? Development and validation of a risk model of in-hospital

mortality after acute myocardial infarction. BMJ Open 2021 Apr 09;11(4):e045053 [FREE Full text] [doi: 10.1136/bmjopen-2020-045053] [Medline: 33837102]
3. Tan Y, Yu F, Long J, Gan L, Wang H, Zhang L, et al. Frequency of systemic lupus erythematosus was decreasing among hospitalized patients from 2013 to 2017 in a national database in China. Front Med (Lausanne) 2021;8:648727 [FREE Full text] [doi: 10.3389/fmed.2021.648727] [Medline: 33889586]
4. Sutherland SM. Electronic health record-enabled big-data approaches to nephrotoxin-associated acute kidney injury risk prediction. Pharmacotherapy 2018 Aug;38(8):804-812. [doi: 10.1002/phar.2150] [Medline: 29885015]
5. Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. The Lancet HIV 2019 Oct;6(10):e688-e695. [doi: 10.1016/s2352-3018(19)30137-7]
6. Chung J, Cho I. The need for academic electronic health record systems in nurse education. Nurse Educ Today 2017 Jul;54:83-88. [doi: 10.1016/j.nedt.2017.04.018] [Medline: 28500984]
7. Shin J, Ko H, Kim JA, Song Y, An JS, Nam SJ, et al. Hospital cancer pain management by electronic health record-based automatic screening. Am J Manag Care 2018 Nov 01;24(11):e338-e343 [FREE Full text] [Medline: 30452201]
8. National Health Commission. 2020 National Medical Services and Quality Safety Report. Beijing: Science and technology literature press; Oct 1, 2021.
9. Notice on printing and distributing the national pilot technical specifications and grouping plans for the payment of disease diagnosis related groups (DRGs). The Central People's Government of the People's Republic of China. URL: https://www.gov.cn/zhengce/zhengceku/2019-11/18/content_5562261.htm [accessed 2023-12-20]
10. Schneider H. Failure mode and effect analysis: FMEA from theory to execution. Technometrics 1996 Feb;38(1):80-80. [doi: 10.1080/00401706.1996.10484424]
11. Fu Y, Qin Y, Wang W, Liu X, Jia L. An extended FMEA model based on cumulative prospect theory and type-2 intuitionistic fuzzy VIKOR for the railway train risk prioritization. Entropy 2020 Dec 15;22(12):1418. [doi: 10.3390/e22121418]
12. Lee JC, Daraba A, Voidarou C, Rozos G, Enshasy HAE, Varzakas T. Implementation of food safety management systems along with other management tools (HAZOP, FMEA, Ishikawa, Pareto). The case study of Listeria monocytogenes and correlation with microbiological criteria. Foods 2021 Sep 13;10(9):2169. [doi: 10.3390/foods10092169]
13. Edu AS, Agoyi M, Agozie D. Digital security vulnerabilities and threats implications for financial institutions deploying digital technology platforms and application: FMEA and FTOPSIS analysis. Peer J Comput Sci 2021;7:e658 [FREE Full text] [doi: 10.7717/peerj-cs.658] [Medline: 34435101]
14. Liu H, Zhang L, Ping Y, Wang L. Failure mode and effects analysis for proactive healthcare risk evaluation: A systematic literature review. J Eval Clin Pract 2020 Aug;26(4):1320-1337. [doi: 10.1111/jep.13317] [Medline: 31849153]
15. Notice of the Office of the National Medical Security Administration on issuing the detailed grouping plan for diagnosis related groups of medical security diseases (CHS-DRG) (version 1.0). The Central People's Government of the People's Republic of China. URL: https://www.gov.cn/zhengce/zhengceku/2019-11/18/content_5562261.htm [accessed 2023-12-20]
16. Asgari Dastjerdi H, Khorasani E, Yarmohammadian MH, Ahmadzade MS. Evaluating the application of failure mode and effects analysis technique in hospital wards: a systematic review. J Inj Violence Res 2017 Jan;9(1):51-60 [FREE Full text] [doi: 10.5249/jivr.v9i1.794] [Medline: 28039688]
17. Harvey HB, Sotardi ST. The Pareto principle. J Am Coll Radiol 2018 Jun;15(6):931. [doi: 10.1016/j.jacr.2018.02.026] [Medline: 29706287]
18. Vázquez L, Salavert M, Gayoso J, Lizasoaín M, Ruiz Camps I, Di Benedetto N, Study Group of Risk Factors for IFI using the Delphi Method. Delphi-based study and analysis of key risk factors for invasive fungal infection in haematological patients. Rev Esp Quimioter 2017 Apr;30(2):103-117 [FREE Full text] [Medline: 28198173]
19. Jones CH, Wylie V, Ford H, Fawell J, Holmer M, Bell K. A robust scenario analysis approach to water recycling quantitative microbial risk assessment. J Appl Microbiol 2023 Mar 01;134(3):029-029. [doi: 10.1093/jambio/lxad029] [Medline: 36796790]
20. Dominguez JA, Pacheco LA, Moratalla E, Carugno JA, Carrera M, Perez-Milan F, et al. Diagnosis and management of isthmocele (Cesarean scar defect): a SWOT analysis. Ultrasound Obstet Gynecol 2023 Sep;62(3):336-344. [doi: 10.1002/uog.26171] [Medline: 36730180]
21. Maughan NM, Garcia-Ramirez JL, Huang FS, Willis DN, Iravani A, Amurao M, et al. Failure modes and effects analysis of pediatric I-131 MIBG therapy: Program design and potential pitfalls. Pediatr Blood Cancer 2022 Dec;69(12):e29996. [doi: 10.1002/pbc.29996] [Medline: 36102748]
22. Anjalee JAL, Rutter V, Samaranayake NR. Application of failure mode and effects analysis (FMEA) to improve medication safety in the dispensing process - a study at a teaching hospital, Sri Lanka. BMC Public Health 2021 Jul 20;21(1):1430 [FREE Full text] [doi: 10.1186/s12889-021-11369-5] [Medline: 34284737]
23. Lin L, Wang R, Chen T, Deng J, Niu Y, Wang M. Failure mode and effects analysis on the control effect of multi-drug-resistant bacteria in ICU patients. Am J Transl Res 2021;13(9):10777-10784 [FREE Full text] [Medline: 34650755]
24. Frosini F, Miniati R, Grillone S, Dori F, Gentili GB, Belardinelli A. Integrated HTA-FMEA/FMECA methodology for the evaluation of robotic system in urology and general surgery. THC 2016 Nov 14;24(6):873-887. [doi: 10.3233/thc-161236]

25. Yang Y, Chang Q, Chen J, Zou X, Xue Q, Song A. Application of integrated emergency care model based on failure modes and effects analysis in patients with ischemic stroke. Front Surg 2022;9:874577 [FREE Full text] [doi: 10.3389/fsurg.2022.874577] [Medline: 35449548]
26. Carlson D, Wallace CJ, East TD, Morris AH. Proc Annu Symp Comput Appl Med Care 1995:188-192 [FREE Full text] [Medline: 8563264]
27. Hart R, Kuo MH. Better data quality for better healthcare research results - a case study. Stud Health Technol Inform 2017;234:161-166. [Medline: 28186034]
28. Andreae MC, Dunham K, Freed GL. Inadequate training in billing and coding as perceived by recent pediatric graduates. Clin Pediatr (Phila) 2009 Nov;48(9):939-944. [doi: 10.1177/0009922809337622] [Medline: 19483135]
29. Balla F, Garwe T, Motghare P, Stamile T, Kim J, Mahnken H, et al. Evaluating coding accuracy in General Surgery Residents' Accreditation Council for Graduate Medical Education procedural case logs. J Surg Educ 2016;73(6):e59-e63. [doi: 10.1016/j.jsurg.2016.07.017] [Medline: 27886974]
30. Greenky MR, Winters BS, Bishop ME, McDonald EL, Rogero RG, Shakked RJ, et al. Coding education in residency and in practice improves accuracy of coding in orthopedic surgery. Orthopedics 2020 Nov 01;43(6):380-383. [doi: 10.3928/01477447-20200827-10] [Medline: 32882048]
31. Glauser G, Sharma N, Beatson N, Dimentberg R, Savarese F, Gagliardi M, et al. Surgical CPT coding discrepancies: analysis of surgeons and employed coders. Am J Med Qual 2021;36(4):263-269. [doi: 10.1177/1062860620959440] [Medline: 32959674]
32. Willyard C. Can AI fix medical records? Nature 2019 Dec;576(7787):S59-S62. [doi: 10.1038/d41586-019-03848-y] [Medline: 31853075]
33. Kilic A. Artificial intelligence and machine learning in cardiovascular health care. Ann Thorac Surg 2020 May;109(5):1323-1329. [doi: 10.1016/j.athoracsur.2019.09.042] [Medline: 31706869]
34. Reinus JF. The electronic medical record, Lawrence Weed, and the quality of clinical documentation. Am J Med 2021 Mar;134(3):e143-e144. [doi: 10.1016/j.amjmed.2020.09.055] [Medline: 33171102]
35. Human factors engineering and user-centered design for mobile health technology: enhancing effectiveness, efficiency, and satisfaction. In: Human-Automation Interaction. New York: Springer, Cham; Dec 15, 2022.
36. 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]
37. Lamovšek N, Klun M. Evaluation of biomedical laboratory performance optimisation using the DEA method. Zdr Varst 2020 Sep;59(3):172-179 [FREE Full text] [doi: 10.2478/sjph-2020-0022] [Medline: 32952718]
38. Alizadeh SS, Solimanzadeh Y, Mousavi S, Safari GH. Risk assessment of physical unit operations of wastewater treatment plant using fuzzy FMEA method: a case study in the northwest of Iran. Environ Monit Assess 2022 Jul 23;194(9):609. [doi: 10.1007/s10661-022-10248-9] [Medline: 35870035]

Abbreviations

EMR: electronic medical record
FMEA: failure mode and effects analysis
RPN: risk priority number
SWOT: strengths, weaknesses, opportunities, and threats

Edited by J Hefner; submitted 21.09.23; peer-reviewed by C Or, Y Zhang, PH Liao; comments to author 24.10.23; revised version received 24.11.23; accepted 05.12.23; published 19.01.24.

Please cite as:
Zhan S, Ding L, Li H, Su A
Application of Failure Mode and Effects Analysis to Improve the Quality of the Front Page of Electronic Medical Records in China: Cross-Sectional Data Mapping Analysis
JMIR Med Inform 2024;12:e53002
URL: https://medinform.jmir.org/2024/1/e53002
doi:10.2196/53002
PMID:38241064

Viewpoint

Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review

Elham Asgari[1,2], MD, MSc, PhD; Japsimar Kaur[3], MBBS; Gani Nuredini[4], MBBS, BMedSci; Jasmine Balloch[2], BSc, MSc; Andrew M Taylor[5], MBBS, Prof Dr; Neil Sebire[5], MBBS, MD, Prof Dr; Robert Robinson[5], MBBS, MA, PhD; Catherine Peters[5], MBChB, MD; Shankar Sridharan[5], BSc, MBBS; Dominic Pimenta[2], MBBS, BSc

[1]Guy's and St Thomas' NHS Trust, London, United Kingdom

[2]Tortus AI, London, United Kingdom

[3]Manchester University NHS Foundation Trust, Manchester, United Kingdom

[4]Barts Health NHS Trust, London, United Kingdom

[5]Great Ormond Street Hospital, London, United Kingdom

Corresponding Author:
Elham Asgari, MD, MSc, PhD
Tortus AI
193-197 High Holborn
London, WC1V 7BD
United Kingdom
Phone: 44 7763891802
Email: asgelham@gmail.com

Abstract

The cognitive load theory suggests that completing a task relies on the interplay between sensory input, working memory, and long-term memory. Cognitive overload occurs when the working memory's limited capacity is exceeded due to excessive information processing. In health care, clinicians face increasing cognitive load as the complexity of patient care has risen, leading to potential burnout. Electronic health records (EHRs) have become a common feature in modern health care, offering improved access to data and the ability to provide better patient care. They have been added to the electronic ecosystem alongside emails and other resources, such as guidelines and literature searches. Concerns have arisen in recent years that despite many benefits, the use of EHRs may lead to cognitive overload, which can impact the performance and well-being of clinicians. We aimed to review the impact of EHR use on cognitive load and how it correlates with physician burnout. Additionally, we wanted to identify potential strategies recommended in the literature that could be implemented to decrease the cognitive burden associated with the use of EHRs, with the goal of reducing clinician burnout. Using a comprehensive literature review on the topic, we have explored the link between EHR use, cognitive load, and burnout among health care professionals. We have also noted key factors that can help reduce EHR-related cognitive load, which may help reduce clinician burnout. The research findings suggest that inadequate efforts to present large amounts of clinical data to users in a manner that allows the user to control the cognitive burden in the EHR and the complexity of the user interfaces, thus adding more "work" to tasks, can lead to cognitive overload and burnout; this calls for strategies to mitigate these effects. Several factors, such as the presentation of information in the EHR, the specialty, the health care setting, and the time spent completing documentation and navigating systems, can contribute to this excess cognitive load and result in burnout. Potential strategies to mitigate this might include improving user interfaces, streamlining information, and reducing documentation burden requirements for clinicians. New technologies may facilitate these strategies. The review highlights the importance of addressing cognitive overload as one of the unintended consequences of EHR adoption and potential strategies for mitigation, identifying gaps in the current literature that require further exploration.

(JMIR Med Inform 2024;12:e55499) doi:10.2196/55499

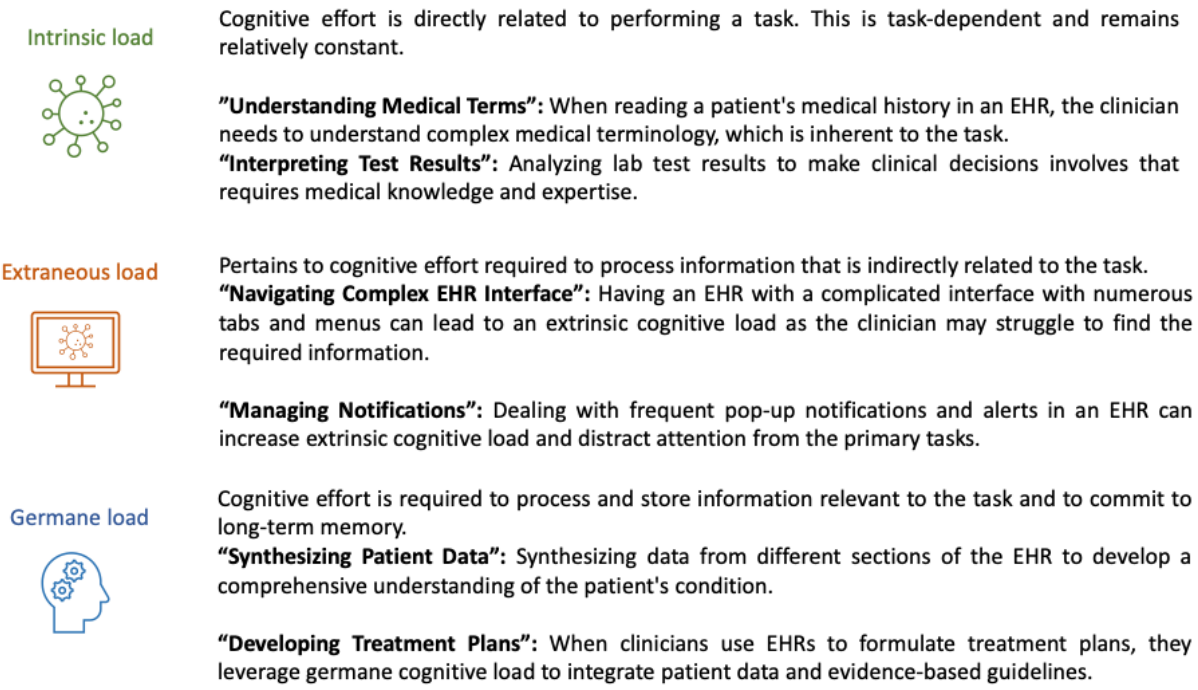
KEYWORDS

electronic health record; cognitive load; burnout; technology; clinician

Introduction

Sweller [1] defined cognitive load as "the amount of mental effort required to process and store information in working memory." The cognitive load theory argues that completing a task requires a complex interplay between sensory inputs, working memory, and long-term memory [1]. The working memory helps interpret the sensory input and then commits processed information into long-term memory. This psychological theory stipulates that while both sensory and long-term memories can handle large volumes of input data, working memory has a comparatively limited capacity and can keep 3 to 5 items in mind at a time [2]. When the amount of information exceeds this given capacity, it leads to cognitive overload. For any given task, there are 3 main factors that contribute to its perceived cognitive load: intrinsic, extraneous, and germane [1]. Figure 1 provides a description of each cognitive load with examples of how they are affected when clinicians use electronic health records (EHRs).

Figure 1. Factors contributing to cognitive load and examples of how they are affected when clinicians use EHRs. EHR: electronic health record.



Physician burnout refers to a state of chronic physical and emotional exhaustion experienced by clinicians, often due to prolonged stress and heavy workloads. It is a growing concern in the medical sector, both in the United Kingdom and globally, and it affects individual health care professionals as well as the health care system as a whole. Key aspects include emotional exhaustion, demoralization, and a reduced sense of accomplishment [3].

Contributing factors to burnout include unremitting workloads, administrative burdens, emotional stress, work-life imbalance, the lack of autonomy, and insufficient support systems. Physicians who work long hours and have high patient loads can experience physical and mental fatigue. Additionally, administrative tasks, paperwork, and EHR requirements add to the workload, causing frustration [4].

According to a report from the World Health Organization, the average life expectancy globally has risen by 6 years, from 66.8 years in 2000 to 73.4 years in 2019 [5]. It is estimated that 1 in 3 adults have multiple long-term conditions [6]. With advancements in science and technology, there are now more diagnostic and treatment options available, which require additional monitoring and follow-up. Consequently, physicians face an increased cognitive load due to the larger volume of information they need to process from complex patients to deliver high-quality care.

Simultaneously, the extraneous load experienced by clinicians is influenced by the presentation of this information. When information is disorganized, unnecessary, or incomplete, clinicians are exposed to a higher extraneous load, placing a greater demand on their working memory [7], which in turn has a knock-on effect on the germane load. This effect is most pronounced among early-career clinicians with a significant amount of new material to learn [7].

Working memory is also attenuated in the presence of physiological or emotional stress [8], which in recent times has become increasingly more common among clinicians following the COVID-19 pandemic, leading to widespread burnout.

Although EHR systems are widely used in health care settings worldwide, there are not enough comprehensive studies evaluating their advantages and disadvantages or examining how they can be improved. In a systematic review, Moy and colleagues [9] aimed to identify the studies on physician and nursing burnout related to using EHR systems. They found that 40% of the 35 studies meeting their inclusion criteria had mentioned clinical burnout. However, they noted a lack of standardized and validated measures to assess the documentation

burden related to EHR use. There is also a lack of objective measures to assess the cognitive load associated with using EHRs.

In summary, the increasing cognitive load experienced by physicians is a sum of the increasing complexity of the information presented, the quality and clarity of that presentation, as well as the emotional and psychological context in which the information is received [1]. The usability of EHRs, including their design, interoperability, and various regulatory requirements, impacts this cognitive burden. By working constantly above the cognitive load threshold, clinicians may exhibit symptoms of cognitive overload, which is considered an immediate precursor to burnout [10]. This narrative review examines the existing literature on cognitive overload experienced by health care professionals, specifically in relation to their use of EHRs and the associated burnout. The review also explores potential solutions that could help reduce the EHR-related cognitive load and improve the well-being of clinicians.

Use of EHRs in Health Care

The digitization of health care has been a growing trend over the past decades, which has seen most patient data transferred from paper records to EHRs. The use of EHR dates back to the 1960s but was only limited to government use [11]. Since the 1970s, EHR systems have been developed with hierarchical or relational databases for various indications, such as to help with hospital billing and scheduling systems, to help improve medical care, and for use in medical research. As computers have become more accessible, larger health care organizations have begun using them to gather patient information [12]. This shift has gradually expanded to encompass nonclinical tasks such as administration, medicolegal work, research, and education [13], which have also increased in demand and complexity over time.

Poor physician acceptance and the lack of incentives, in addition to high costs and errors associated with data entry, hindered EHR implementation uptake in the 1990s, and thus, digital records were not widespread [14]. By the 2000s, countries such as the United States and the United Kingdom began implementing national projects to digitize paper records [12]. The UK government attempted to create the National Programme for Information Technology [15] in 2002 to create a universal EHR system for the entire United Kingdom. The nationwide initiative was centered around 3 main objectives: lifelong EHRs, 24/7 web-based access by public health care professionals, and seamless information sharing throughout all sectors of the National Health Service [16]. The project failed to meet its objectives, and digitization became fragmented and slow once again [17]. However, by 2022, a total of 86% of the UK hospital trusts had successfully transitioned from paper notes to a digital system (although only a minority have enterprise-wide EHR capability), with the figure expected to reach 90% by December 2023 [18]. In the United States, the *Health Information Technology for Economic and Clinical Health Act* was signed into law in February 2009 to promote the adoption of and meaningful use of health information technology (HIT) as part of the American Recovery and Reinvestment Act. Financial incentives that were allocated as part of the *Health Information Technology for Economic and Clinical Health Act* led to a significant increase in the adoption of EHRs in the United States [14].

Factors Impacting the Usability of EHR Systems

The usability of EHR systems continues to be a major concern, whereby clinicians are subjected to too much or too little information, preprogrammed workflows, and multiple alerts [19]. There have been problems with chaotic, nonintuitive visual displays and numerous default settings that might not be relevant for a given task or patient [20]. Navigating through the same information includes unnecessary steps, for example, multiple clicks and duplicated information [21]. Users experience higher fatigue, leading to potential room for errors and decreased efficiency [22,23]. In addition to documentation and chart review, managing inbox tasks has been noted as one of the significant burdens for clinicians [24]. Receiving excessive notification has been shown to cause alert fatigue, leading to missing important information and poor patient outcomes [25]. Although clinical decision support systems have been introduced to enhance patient care, excessive use of interruptive clinical decision support systems in the EHRs can lead to alert fatigue and reduced effectiveness. Chaparro and colleagues [26] have described how interruptive alerts can increase cognitive burden and lead to reduced acceptance of the alert and an increase in the number of errors.

Several factors have been highlighted as contributing to the use of EHR and physician well-being. Nguyen and colleagues [27] studied this in a systematic review, where they found that EHR-related physician well-being is determined by multiple factors, including EHR usability, EHR system features, and physician-level characteristics and beliefs.

The sheer volume of data that a physician can access during a specific clinical encounter proves challenging [28]. As an example, Hill and colleagues [29] found that emergency health care physicians see an average of 2.4 patients per hour and use 4000 mouse clicks in a 10-hour shift. This can result from a combination of poor EHR design and information overload and adds to physician stress [30,31]. Information overload is a part of the 5 main hazards of “information chaos” alongside information underload, information scatter, information conflict, and erroneous information [32]. Clinicians are then required to spend more effort to filter through the information, clarify conflicting documentation, or reassess potentially erroneous information, leading to excess workload and adverse outcomes on not only patient care and health systems but, more importantly, clinician well-being [33].

Gal and colleagues [34] studied this in a pediatric intensive care unit where they calculated that each patient generated an average of 1460 new data points in a 24-hour period. Pediatric intensive care unit attending physicians cared for an average of 11 patients during the day and 22 patients overnight, resulting in exposure to 16,060 (range 11,680-18,980) and 32,120 (range

23,360-37,960) individual data points during the day and night, respectively.

Wanderer and colleagues [35] have described how optimal data visualization in various specialties can lead to improved decision-making for clinicians and more efficient use of their time. Many EHR vendors use visual analytic systems to improve physician workflow and reduce medical errors [36].

Blink rate, measured using eye-tracking technology, has been associated with cognitive workload. Visual tasks that require more focused attention and working memory load have been shown to reduce blink rate [37]. A decreased blink rate has been found to occur in EHR-based tasks that require more cognitive workload [38].

The NASA Task-Load Index (NASA-TLX) is a widely used questionnaire to assess perceived workload (available in Multimedia Appendix 1) [39]. It consists of 6 questions, which can be rated from 1 to 10. Nurses rated their perceived workload from 0 (very low) to 10 (very high).

Using blink rate in addition to the NASA-TLX, Mazur and colleagues [40] tested the implications of the EHR usability interface in a study where they assigned tasks, including the review of medical test results for 20 and 18 individuals using baseline and enhanced EHR versions, respectively, that provided policy-based decision support instructions for next steps. Interestingly, they found that the baseline group had poorer performance and higher cognitive load compared with those who used the enhanced version, suggesting the importance of improving the usability of EHRs to address issues such as clinician burnout and patient safety events.

Harry and colleagues [7] studied the direct relationship between cognitive load with physician burnout in a national sample of US physicians. Using the NASA-TLX, they had responses from 4517 (85.6%) of the 5276 physicians included in the survey. The median age of the physicians was 53 years; 61.8% were male, 37.9% were female, and 0.3% were other gender; and 24 specialties were identified. They identified a dose-response relationship between physician task load and the risk of burnout independent of age, gender, practice setting, and hours worked per week.

To demonstrate a more accurate association between EHR use and stress, Yen and colleagues [41] used blood pulse wave monitoring (previously used as a surrogate for chronic stress) in addition to NASA-TLX on 7 nurses during 132 hours of work. They found that the nursing staff spent 45.54 minutes using EHR during a 4-hour shift, which was much more than the time spent on any other communication or hands-on activities. In addition, the nurses' stress when using EHR was associated with higher perceived physical demand and frustration.

The level of EHR-related burnout has also been shown to be in part influenced by physician specialty. In a large study that used assessing questionnaires among physicians in various specialties with over 25,000 respondents, the investigators found the level of burnout ranged from 22% to 34% by specialty [42]. The specialties with the highest levels of burnout were family medicine (34%) and hematology or oncology (33%). The specialties with the lowest levels of burnout were psychiatry (22%) and anesthesiology (24%). After adjusting for confounding variables, physicians with 5 or fewer hours of weekly out-of-hours charting were twice as likely to report lower levels of burnout than those with 6 or more hours. Those who agree that their organization has performed well with EHR implementation, training, and support were also twice as likely to report lower levels of burnout than those who disagreed. This highlights the importance of training and support following the implementation of EHR for their optimal use.

In a scoping review, Muhiyaddin and colleagues [43] studied the causes and consequences of physician burnout related to the use of EHRs. Reviewing 30 eligible studies out of 500, they identified 6 main causes that are related to physician burnout, including EHR documentation and related tasks, poor design of EHR systems, workload leading to overtime work, inbox alerts, and alert fatigue. Not surprisingly, physician burnout was associated with a low quality of care, behavioral issues, and mental health complications, as well as career dissatisfaction and a reduction in patient safety and satisfaction.

In a survey of 640 clinicians from 3 institutions, with 282 (44.1%) responses to 105 questions, Kroth and colleagues [30] identified 7 EHR design and use factors associated with high stress and burnout. These were information overload, slow system response times, excessive data entry, inability to navigate the system quickly, note bloat, interference with the patient-clinician relationship, fear of missing something, and notes geared toward billing.

Another study [44] aiming to quantify burnout due to the use of HIT used a survey sent to 4197 physicians, where 1792 responded (response rate: 42.7%). They found that HIT-related stress was measurable, prevalent, and specialty related. About 70% of physicians with EHRs experienced HIT-related stress in their sample, and the presence of any of the 3 HIT-related stress measures independently predicted burnout symptoms among respondents. In particular, those with time pressures for documentation or those doing excessive "work after work" on their EHR at home had approximately twice the odds of burnout compared to physicians without these challenges. Time spent after hours on the EHR and the volume of inbox messages have been found to relate to physician exhaustion [45].

Using live observational design and NASA-TLX surveys, Khairat and colleagues [46] assessed the effect of EHRs on emergency department attending and resident physicians' perceived workload, satisfaction, and productivity through completing 6 EHR patient scenarios. They found that EHR frustration levels are significantly higher among more senior attending physicians compared with more junior resident physicians. Among the factors causing high EHR frustrations are (1) remembering menu and button names and commands use; (2) performing tasks that are not straightforward; (3) system speed; and (4) system reliability.

Advantages and Disadvantages of Using EHRs

Overview

As highlighted in the previous section, despite their potential benefits, there have been growing concerns that EHRs also have detrimental effects. Here, we summarize some of the advantages and disadvantages of using EHRs.

Information overload is a significant concern when using EHRs [47-49]. Various studies also suggest a correlation between the usability of the EHR and cognitive load and burnout among clinicians [34,50-52]. Clinicians feel that work-life balance, satisfaction rates, attrition, and burnout are all affected due to the continuous daily interaction with EHR systems [22,53-55].

Advantages

The transition from paper-based medical records to EHRs has been perceived as a positive development in several areas [9,56]. In addition to being easily accessible, EHR systems have been shown to improve communication between clinicians and enhance the continuity of care [57,58]. They can lead to better-informed decisions due to the availability of data and avoid the duplication of diagnostic testing [59]. However, a review of the impact of EHR use on enhancing medication safety, one of the biggest risks to patient care, has shown only modest improvements [60].

EHRs also contain a high volume of clinical data, providing us with multiple opportunities to conduct research and audit [13,59]. A good example of this in the United Kingdom is OpenSAFELY, a secure, transparent, and open-source software platform for the analysis of EHR data [61]. During the COVID-19 pandemic, scientists and statisticians could use the data available on this platform to provide insights into population demographics most at risk of death following COVID-19 infection, which aided with the national policy strategy for prioritizing care [62-65].

Disadvantages

Over the past decade, there has been a reported increase in burnout levels among clinicians, with one potential factor being the introduction of EHR systems [23,34]. The introduction of EHRs has resulted in changes in workflow, with frontline clinicians taking on administrative tasks such as ordering tests, correcting notes, and placing referrals. This has led to increased cognitive load, which is often overlooked [66,67].

On a day-to-day basis, clinicians face time constraints; administrative load; and consequently, elongated workdays. The current documentation methods used in EHRs are under scrutiny by clinicians due to the perceived poor quality of user interfaces, ultimately leading to burnout [52]. Factors such as increased structured documentation requirements, physician order entry, inbox management, and patient portals contribute to more work that is not direct face time with patients [19,68].

Inflated documentation also extends to the excessive use of templates and copy-and-paste workflows in EHR systems that introduce data that are neither required nor accurate [48,49,69]. EHRs allow information to be copied from almost anywhere in the record to another section. This can save time and allow clinicians to focus on clinical tasks rather than documentation; however, it comes with its own challenges. Erroneous information can be copied and pasted without editing, leading to data integrity issues and diagnostic errors [70]. This also creates room for false assumptions and inferred incorrect information between different health care professions, perpetuating previous inaccuracies. It might also sanction junior clinicians to rely solely on readily available information rather than conducting a thorough history and examination for themselves and constructing their own differential thought processes [71].

Although thorough documentation is key for clinical care, there has been a rise in complex and lengthy documentation of content that is required for billing purposes, quality improvement measures, avoiding malpractice, and signs of compliance [72]. In countries such as the United States, the regulatory requirements for data entry beyond what is required for direct patient care can contribute to an increasing workload [73]. Examples of these include collecting data for claim submission, prior authorization, billing, and quality reporting. In addition, a lack of interoperability between EHR systems can result in clinicians not having access to adequate patient information and fragmented care [74]. Often the clinical needs to spend a significant amount of time to obtain this information from various medical records between different health care organizations and sometimes even within one facility. Textbox 1 summarizes some of the advantages and disadvantages of using EHRs.

Textbox 1. Advantages and disadvantages of using electronic health records (EHRs).

Advantages of using EHRs

- Improved communication between clinicians
- Remote access to clinical records enhances care delivery
- Convenient access to patient information for clinicians
- Facilitates research and audit through a high volume of clinical data storage

Disadvantages of using EHRs

- Information overload leading to cognitive overload
- Increased cognitive load due to EHRs can contribute to feelings of exhaustion and burnout
- Continuous interaction with EHRs affects work-life balance and may lead to burnout
- Complex and lengthy documentation required for billing and quality reporting can be cumbersome
- Poor quality of user interfaces in EHRs leads to clinician burnout
- Excessive use of templates and copy-and-paste workflows can lead to data integrity issues
- The lack of interoperability between EHR systems can lead to missed information and duplication of investigations

Overcoming EHR-Related Burnout

Health care organizations and policy makers worldwide are increasingly recognizing the importance of addressing clinician burnout [75]. Here, we have summarized some of the interventions recommended in the literature that can reduce various types of cognitive load and potentially clinician burnout related to EHR use.

Dymek and colleagues [24] have made a case for producing an evidence base to reduce EHR-related clinician burden. Describing documentation, chart review, and inbox tasks as some of the key contributing factors causing burnout, they have made suggestions to help overcome these challenges. Some of these approaches include using speech recognition software and natural language processing to help with documentation and the generation of progress notes; the use of natural language processing and machine learning to process, filter, and rank patient information so that the attention can be paid to where it is most needed; and the use of better inbox design by involving clinicians in their development. Understanding the workflow of the clinicians and involving them in the design of the EHR have been shown to positively impact its usability and user satisfaction [76].

Several studies have reviewed alert burden in EHRs and described potential solutions on how to manage them effectively. One of the very interesting and useful recommended suggestions is developing an Interruptive Alert Stewardship by implementing metrics to assess the alert burden and their effectiveness in improving outcomes [26]. McGreevey and colleagues [77] have comprehensively described reasons for alert fatigue and suggest that organizations develop an alert governance specific to their needs. They propose that key stakeholders including clinicians, informatics, information technology, and administration groups need to participate in developing the alert governance and oversee the design and purpose of alerts. They also recommend using a checklist to assess the purpose and justification of alerts and suggest using metrics to assess their effectiveness and efficiency. Organizations such as Geisinger Health System and Penn Medicine have successfully improved their EHR alert to help with clinician well-being [77].

Clinicians have specific and feasible suggestions for reducing EHR-related burdens, such as providing high-quality EHR training; having an on-site EHR support team; involving support staff or scribes in the documentation process; and, importantly, obtaining physician input and feedback in improving EHRs [27]. Future efforts should focus on implementing the strategies and upgrades requested by these frontline users.

In a recent systematic review, Kang and Sarkar [78] looked at interventions that have been used to reduce EHR-related burnout. The study identified 3 primary interventions, including the use of scribes, EHR training, EHR modifications, and a combination of training and modifications. The use of scribes has been overall well received by clinicians and patients and, in some cases, led to increased productivity, but there were downsides, in particular, the cost, which would be difficult to overcome in smaller centers. EHR training had varying outcomes, with some studies showing a reduction in documentation time, whereas others did not demonstrate this benefit. Nevertheless, subjective EHR proficiency increased, which could help improve clinicians' perception of EHR.

The study [78] has also examined several EHR modification techniques, such as data entry automation technology, improving EHR workflows, reducing unnecessary inbox alerts, and providing support teams to resolve EHR issues promptly. These interventions resulted in positive outcomes, such as a reduction in documentation time ranging from 18.5 to 60%, improved documentation quality and completion rate, decrease in data errors, and subjective EHR usability and satisfaction. However, these positive effects did not lead to a significant reduction in physician burnout. The authors suggest that this could be due to the fact that although EHR enhancements can improve some aspects of the clinician's workflow, they probably do not address the defects in the EHR usability, which contribute to burnout. In addition, there are other factors contributing to burnout, such

as overall workload, organizational culture, and work-life balance that extend beyond EHR systems.

Improving interoperability and health information exchange through understanding the barriers, appropriate incentives, and legislation can facilitate the clinicians' workflow, reduce workload, and enhance patient safety [79].

In 2020, The Office of the National Coordinator for Health Information Technology published a report outlining strategies to reduce regulatory and administrative burden related to the use of HIT and EHR systems [80]. The report focuses on the challenges of EHR and HIT-related burden, which hinder the achievement of interoperability. They recognize that these burdens increase the time and expense clinicians must invest in interacting with EHRs, reducing the value of information, and diverting resources from patient care. They propose a framework for trusted exchange among health information networks to reduce clinician burden while benefiting patients and the health care system.

The National Academy of Medicine published a potential roadmap for EHR optimization and clinician well-being [81]. They have described several strategies currently available that can improve the usability of EHRs, such as EHR optimization, in-basket management techniques, documentation strategies, team-based workflow, and EHR training, as well as the use of artificial intelligence and add-on applications that can help with interoperability, automation, and decision support tools in the future. Gandhi and colleagues [82] have described how the use of artificial intelligence can reduce the cognitive workload by helping with data gathering, documentation, and decision support. They also suggest useful methods to assess the impact of these technologies.

Gaps in the Current Literature

Given the increasingly interconnected digital ecosystem and the complexity of health care systems, which are influenced by physical, emotional, and human factors, it can be challenging to attribute specific outcomes to any particular technology.

Therefore, to maximize the benefits of new tools, it is essential to create frameworks for scientifically assessing the impact of any technology used.

By using user-friendly interfaces, customization options, and context-sensitive information presentation, EHRs can streamline data management [83]. Incorporating decision support tools, data visualization techniques, and smart documentation practices further enhances health care staff's ability to focus on patient care, reducing the risk of errors and burnout [36]. EHR optimization to support clinical workflow and real-life working is a key to uplifting the well-being of health care professionals.

Addressing physician burnout requires systemic changes, including improving work environments with a renewed focus on teamwork, reducing administrative burdens, providing support, and promoting work-life balance within health care organizations.

Individual strategies, such as self-care, stress management, and professional pastoral help, are crucial for clinicians to mitigate and recover from burnout [84]. This will support the well-being of the health care workforce and ensure ongoing high-quality care delivery for all.

As our patients and work environments become more complex and more health technology products become available, it is crucial that we assess their impact through studies and engaging with our health care staff and patients throughout all stages of their development and use [85].

Conclusion

The use of EHR systems may provide benefit for centralizing patient data and simplifying the process of reviewing records, requesting laboratory and imaging tests, and reviewing results, as well as conducting clinical audits, research, and quality improvement projects.

However, there is a noticeable difference in the quality of various EHR systems health care organizations use. Many of these EHR systems do not communicate with each other, keeping data isolated in silos. Our review highlights the cognitive load that their use places on clinical staff, which is not always considered. Improving the design, user interface, and data visualization or retrieval of EHR systems can help to reduce cognitive load, support working memory, and potentially reduce physician workload while enhancing patient care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
NASA Task-Load Index questionnaire.
[DOCX File , 14 KB - medinform_v12i1e55499_app1.docx]

References

1. Sweller J. Cognitive load during problem solvingffects on learning. Cogn Sci 1988 Apr;12(2):257-285. [doi: 10.1016/0364-0213(88)90023-7]
2. Cowan N. The magical mystery four: how is working memory capacity limited, and why? Curr Dir Psychol Sci 2010 Feb 01;19(1):51-57 [FREE Full text] [doi: 10.1177/0963721409359277] [Medline: 20445769]

3. McKinley N, McCain RS, Convie L, Clarke M, Dempster M, Campbell WJ, et al. Resilience, burnout and coping mechanisms in UK doctors: a cross-sectional study. BMJ Open 2020 Jan 27;10(1):e031765 [FREE Full text] [doi: 10.1136/bmjopen-2019-031765] [Medline: 31988223]
4. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. J Intern Med 2018 Jun;283(6):516-529 [FREE Full text] [doi: 10.1111/joim.12752] [Medline: 29505159]
5. Department of Data and Analytics: Division of Data, Analytics and Delivery for Impact. GHE: life expectancy and healthy life expectancy. World Health Organization. 2020. URL: https://tinyurl.com/4zc9csw4 [accessed 2024-04-05]
6. Hajat C, Stein E. The global burden of multiple chronic conditions: narrative review. Prev Med Rep 2018 Dec;12:284-293 [FREE Full text] [doi: 10.1016/j.pmedr.2018.10.008] [Medline: 30406006]
7. Harry E, Sinsky C, Dyrbye LN, Makowski MS, Trockel M, Tutty M, et al. Physician task load and the risk of burnout among US physicians in a national survey. Jt Comm J Qual Patient Saf 2021 Feb;47(2):76-85 [FREE Full text] [doi: 10.1016/j.jcjq.2020.09.011] [Medline: 33168367]
8. Chajut E, Algom D. Selective attention improves under stress: implications for theories of social cognition. J Pers Soc Psychol 2003 Aug;85(2):231-248 [FREE Full text] [doi: 10.1037/0022-3514.85.2.231] [Medline: 12916567]
9. Moy AJ, Schwartz JM, Chen R, Sadri S, Lucas E, Cato KD, et al. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. J Am Med Inform Assoc 2021 Apr 23;28(5):998-1008 [FREE Full text] [doi: 10.1093/jamia/ocaa325] [Medline: 33434273]
10. Iskander M. Burnout, cognitive overload, and metacognition in medicine. Med Sci Educ 2019 Mar;29(1):325-328 [FREE Full text] [doi: 10.1007/s40670-018-00654-5] [Medline: 34457483]
11. Stone CP. A glimpse at EHR implementation around the world: the lessons the US can learn. Dokumen. 2014 May. URL: https://dokumen.tips/download/link/a-glimpse-at-ehr-implementation-around-the-world-the-glimpse-at-ehr-implementation.html [accessed 2024-04-05]
12. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016 May 20;Suppl 1(Suppl 1):S48-S61 [FREE Full text] [doi: 10.15265/IYS-2016-s006] [Medline: 27199197]
13. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. J Intern Med 2013 Dec 18;274(6):547-560 [FREE Full text] [doi: 10.1111/joim.12119] [Medline: 23952476]
14. Jha AK, DesRoches CM, Kralovec PD, Joshi MS. A progress report on electronic health records in U.S. hospitals. Health Aff (Millwood) 2010 Oct;29(10):1951-1957 [FREE Full text] [doi: 10.1377/hlthaff.2010.0502] [Medline: 20798168]
15. Crompton P. The National Programme for Information Technology--an overview. J Vis Commun Med 2007 Jun;30(2):72-77 [FREE Full text] [doi: 10.1080/17453050701496334] [Medline: 17671907]
16. The electronic health records system in the UK. Centre for Public Impact. URL: https://www.centreforpublicimpact.org/case-study/electronic-health-records-system-uk [accessed 2024-04-05]
17. Justinia T. The UK's National Programme for IT: why was it dismantled? Health Serv Manage Res 2017 Feb;30(1):2-9 [FREE Full text] [doi: 10.1177/0951484816662492] [Medline: 28166675]
18. A plan for digital health and social care. NHS England. URL: https://tinyurl.com/cp2p2w4m [accessed 2022-06-29]
19. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
20. Williams MS. Misdiagnosis: burnout, moral injury, and implications for the electronic health record. J Am Med Inform Assoc 2021 Apr 23;28(5):1047-1050 [FREE Full text] [doi: 10.1093/jamia/ocaa244] [Medline: 33164089]
21. Bouamrane MM, Mair FS. A study of general practitioners' perspectives on electronic medical records systems in NHSScotland. BMC Med Inform Decis Mak 2013 May 21;13:58 [FREE Full text] [doi: 10.1186/1472-6947-13-58] [Medline: 23688255]
22. Babbott S, Manwell LB, Brown R, Montague E, Williams E, Schwartz M, et al. Electronic medical records and physician stress in primary care: results from the MEMO Study. J Am Med Inform Assoc 2014 Feb;21(e1):e100-e106 [FREE Full text] [doi: 10.1136/amiajnl-2013-001875] [Medline: 24005796]
23. Khairat S, Coleman C, Ottmar P, Bice T, Carson SS. Evaluation of physicians' electronic health records experience using actual and perceived measures. Perspect Health Inf Manag 2022 Jan 1;19(1):1k [FREE Full text] [Medline: 35440931]
24. 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 Apr 23;28(5):1057-1061 [FREE Full text] [doi: 10.1093/jamia/ocaa238] [Medline: 33340326]
25. Powell L, Sittig DF, Chrouser K, Singh H. Assessment of health information technology-related outpatient diagnostic delays in the US Veterans Affairs health care system: a qualitative study of aggregated root cause analysis data. JAMA Netw Open 2020 Jun 01;3(6):e206752 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.6752] [Medline: 32584406]
26. 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 [FREE Full text] [doi: 10.1055/s-0042-1748856] [Medline: 35613913]

27. Nguyen OT, Jenkins NJ, Khanna N, Shah S, Gartland AJ, Turner K, et al. A systematic review of contributing factors of and solutions to electronic health record-related impacts on physician well-being. J Am Med Inform Assoc 2021 Apr 23;28(5):974-984 [FREE Full text] [doi: 10.1093/jamia/ocaa339] [Medline: 33517382]
28. 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]
29. Hill RG, Sears LM, Melanson SW. 4000 Clicks: a productivity analysis of electronic medical records in a community hospital ED. Am J Emerg Med 2013 Nov;31(11):1591-1594 [FREE Full text] [doi: 10.1016/j.ajem.2013.06.028] [Medline: 24060331]
30. Kroth PJ, Morioka-Douglas N, Veres S, Babbott S, Poplau S, Qeadan F, et al. Association of electronic health record design and use factors with clinician stress and burnout. JAMA Netw Open 2019 Aug 02;2(8):e199609 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.9609] [Medline: 31418810]
31. Shoolin J, Ozeran L, Hamann C, Bria W. Association of Medical Directors of Information Systems consensus on inpatient electronic health record documentation. Appl Clin Inform 2013 Jun 26;4(2):293-303 [FREE Full text] [doi: 10.4338/ACI-2013-02-R-0012] [Medline: 23874365]
32. Beasley JW, Wetterneck TB, Temte J, Lapin JA, Smith P, Rivera-Rodriguez AJ, et al. Information chaos in primary care: implications for physician performance and patient safety. J Am Board Fam Med 2011;24(6):745-751 [FREE Full text] [doi: 10.3122/jabfm.2011.06.100255] [Medline: 22086819]
33. Patel RS, Bachu R, Adikey A, Malik M, Shah M. Factors related to physician burnout and its consequences: a review. Behav Sci (Basel) 2018 Oct 25;8(11):98 [FREE Full text] [doi: 10.3390/bs8110098] [Medline: 30366419]
34. Gal DB, Han B, Longhurst C, Scheinker D, Shin AY. Quantifying electronic health record data: a potential risk for cognitive overload. Hosp Pediatr 2021 Feb;11(2):175-178 [FREE Full text] [doi: 10.1542/hpeds.2020-002402] [Medline: 33500357]
35. Wanderer JP, Nelson SE, Ehrenfeld JM, Monahan S, Park S. Clinical data visualization: the current state and future needs. J Med Syst 2016 Dec;40(12):275 [FREE Full text] [doi: 10.1007/s10916-016-0643-x] [Medline: 27787779]
36. Rostamzadeh N, Abdullah SS, Sedig K. Visual analytics for electronic health records: a review. Informatics 2021 Feb 23;8(1):12 [FREE Full text] [doi: 10.3390/informatics8010012]
37. Chen S, Epps J. Using task-induced pupil diameter and blink rate to infer cognitive load. Hum Comput Interact 2014 Apr 29;29(4):390-413 [FREE Full text] [doi: 10.1080/07370024.2014.892428]
38. Mosaly PR, Mazur LM, Yu F, Guo H, Derek M, Laidlaw DH, et al. Relating task demand, mental effort and task difficulty with physicians' performance during interactions with electronic health records (EHRs). Int J Hum Comput Interact 2017 Sep 25;34(5):467-475 [FREE Full text] [doi: 10.1080/10447318.2017.1365459]
39. Hart SG. NASA-Task Load Index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2016 Nov 05;50(9):904-908. [doi: 10.1177/154193120605000909]
40. Mazur LM, Mosaly PR, Moore C, Marks L. Association of the usability of electronic health records with cognitive workload and performance levels among physicians. JAMA Netw Open 2019 Apr 05;2(4):e191709 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.1709] [Medline: 30951160]
41. Yen PY, Pearl N, Jethro C, Cooney E, McNeil B, Chen L, et al. Nurses' stress associated with nursing activities and electronic health records: data triangulation from continuous stress monitoring, perceived workload, and a time motion study. AMIA Annu Symp Proc 2019 Mar 4;2019:952-961 [FREE Full text] [Medline: 32308892]
42. Eschenroeder HC, Manzione LC, Adler-Milstein J, Bice C, Cash R, Duda C, et al. Associations of physician burnout with organizational electronic health record support and after-hours charting. J Am Med Inform Assoc 2021 Apr 23;28(5):960-966 [FREE Full text] [doi: 10.1093/jamia/ocab053] [Medline: 33880534]
43. Muhiyaddin R, Elfadl A, Mohamed E, Shah Z, Alam T, Abd-Alrazaq A, et al. Electronic health records and physician burnout: a scoping review. Stud Health Technol Inform 2022 Jan 14;289:481-484. [doi: 10.3233/SHTI210962] [Medline: 35062195]
44. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
45. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 01;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
46. Khairat S, Burke G, Archambault H, Schwartz T, Larson J, Ratwani RM. Perceived burden of EHRs on physicians at different stages of their career. Appl Clin Inform 2018 Apr;9(2):336-347 [FREE Full text] [doi: 10.1055/s-0038-1648222] [Medline: 29768634]
47. What is "cognitive load"—and how can we help clinicians manage it? Nuance. 2022 Aug 11. URL: https://whatsnext.nuance.com/healthcare-ai/cognitive-load-and-impact-on-clinician-burnout/ [accessed 2024-04-05]
48. Downing N, Bates DW, Longhurst CA. Physician burnout in the electronic health record era: are we ignoring the real cause? Ann Intern Med 2018 Jul 03;169(1):50-51. [doi: 10.7326/M18-0139] [Medline: 29801050]

49. Pickering BW, Gajic O, Ahmed A, Herasevich V, Keegan MT. Data utilization for medical decision making at the time of patient admission to ICU. Crit Care Med 2013 Jun;41(6):1502-1510. [doi: 10.1097/CCM.0b013e318287f0c0] [Medline: 23528804]
50. Gawande A. Why doctors hate their computers. New Yorker. 2018 Nov 5. URL: https://www.newyorker.com/magazine/2018/11/12/why-doctors-hate-their-computers [accessed 2024-04-05]
51. Ash JS, Sittig DF, Dykstra RH, Guappone K, Carpenter JD, Seshadri V. Categorizing the unintended sociotechnical consequences of computerized provider order entry. Int J Med Inform 2007 Jun;76 Suppl 1:S21-S27 [FREE Full text] [doi: 10.1016/j.ijmedinf.2006.05.017] [Medline: 16793330]
52. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. J Am Med Inform Assoc 2004;11(2):104-112 [FREE Full text] [doi: 10.1197/jamia.M1471] [Medline: 14633936]
53. Friedberg MW, Chen PG, Van Busum KR, Aunon F, Pham C, Caloyeras J, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Rand Health Q 2014 Dec 1;3(4):1 [FREE Full text] [Medline: 28083306]
54. Shanafelt TD, Hasan O, Dyrbye LN, Sinsky C, Satele D, Sloan J, et al. Changes in burnout and satisfaction with work-life balance in physicians and the general US working population between 2011 and 2014. Mayo Clin Proc 2015 Dec;90(12):1600-2613 [FREE Full text] [doi: 10.1016/j.mayocp.2015.08.023] [Medline: 26653297]
55. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, et al. Burnout and satisfaction with work-life balance among US physicians relative to the general US population. Arch Intern Med 2012 Oct 08;172(18):1377-1385 [FREE Full text] [doi: 10.1001/archinternmed.2012.3199] [Medline: 22911330]
56. Aziz F, Talhelm L, Keefer J, Krawiec C. Vascular surgery residents spend one fifth of their time on electronic health records after duty hours. J Vasc Surg 2019 May;69(5):1574-1579 [FREE Full text] [doi: 10.1016/j.jvs.2018.08.173] [Medline: 31010521]
57. Ball C, McBeth PB. The impact of documentation burden on patient care and surgeon satisfaction. Can J Surg 2021 Aug 10;64(4):E457-E458 [FREE Full text] [doi: 10.1503/cjs.013921] [Medline: 34388108]
58. Alobo IG, Soyannwo T, Ukponwan G, Akogu S, Akpa AM, Ayankola K. Implementing electronic health system in Nigeria: perspective assessment in a specialist hospital. Afr Health Sci 2020 Jun;20(2):948-954 [FREE Full text] [doi: 10.4314/ahs.v20i2.50] [Medline: 33163063]
59. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55 [FREE Full text] [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
60. Ratwani RM. Modest progress on the path to electronic health record medication safety. JAMA Netw Open 2020 May 01;3(5):e206665 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.6665] [Medline: 32469409]
61. Secure analytics platform for NHS electronic health records. OpenSAFELY. URL: https://www.opensafely.org/ [accessed 2024-04-05]
62. Bhaskaran K, Bacon S, Evans SJW, Bates CJ, Rentsch CT, MacKenna B, et al. Factors associated with deaths due to COVID-19 versus other causes: population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. Lancet Reg Health Eur 2021 Jul;6:100109 [FREE Full text] [doi: 10.1016/j.lanepe.2021.100109] [Medline: 33997835]
63. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature 2020 Aug;584(7821):430-436 [FREE Full text] [doi: 10.1038/s41586-020-2521-4] [Medline: 32640463]
64. Issitt RW, Booth J, Bryant WA, Spiridou A, Taylor AM, du Pré P, et al. Children with COVID-19 at a specialist centre: initial experience and outcome. The Lancet Child & Adolescent Health 2020 Aug;4(8):e30-e31 [FREE Full text] [doi: 10.1016/s2352-4642(20)30204-2]
65. Bourgeois FT, Gutiérrez-Sacristán A, Keller MS, Liu M, Hong C, Bonzel CL, et al. International analysis of electronic health records of children and youth hospitalized with COVID-19 Infection in 6 countries. JAMA Netw Open 2021 Jun 01;4(6):e2112596 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.12596] [Medline: 34115127]
66. Colligan L, Potts HWW, Finn CT, Sinkin RA. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. Int J Med Inform 2015 Jul;84(7):469-476 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.03.003] [Medline: 25868807]
67. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB. A model of burnout and life satisfaction amongst nurses. J Adv Nurs 2000 Aug;32(2):454-464 [FREE Full text] [doi: 10.1046/j.1365-2648.2000.01496.x] [Medline: 10964195]
68. Baumann L, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: a systematic review. Health Policy 2018 Aug;122(8):827-836 [FREE Full text] [doi: 10.1016/j.healthpol.2018.05.014] [Medline: 29895467]
69. Tsou A, Lehmann C, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Appl Clin Inform 2017 Dec 20;26(01):12-34 [FREE Full text] [doi: 10.4338/aci-2016-09-r-0150]
70. Vogel L. Cut-and-paste clinical notes confuse care, say US internists. CMAJ 2013 Dec 10;185(18):E826 [FREE Full text] [doi: 10.1503/cmaj.109-4656] [Medline: 24218539]

71. Cheng CG, Wu DC, Lu JC, Yu CP, Lin HL, Wang MC, et al. Restricted use of copy and paste in electronic health records potentially improves healthcare quality. Medicine (Baltimore) 2022 Jan 28;101(4):e28644 [FREE Full text] [doi: 10.1097/MD.0000000000028644] [Medline: 35089204]
72. Koopman RJ, Steege LMB, Moore JL, Clarke MA, Canfield SM, Kim MS, et al. Physician information needs and electronic health records (EHRs): time to reengineer the clinic note. J Am Board Fam Med 2015;28(3):316-323 [FREE Full text] [doi: 10.3122/jabfm.2015.03.140244] [Medline: 25957364]
73. Tutty M, Carlasare LE, Lloyd S, Sinsky CA. The complex case of EHRs: examining the factors impacting the EHR user experience. J Am Med Inform Assoc 2019 Jul 01;26(7):673-677 [FREE Full text] [doi: 10.1093/jamia/ocz021] [Medline: 30938754]
74. Jacob JA. On the road to interoperability, public and private organizations work to connect health care data. JAMA 2015 Sep;314(12):1213-1215 [FREE Full text] [doi: 10.1001/jama.2015.5930] [Medline: 26393833]
75. National Academies of Sciences, Engineering, and Medicine, National Academy of Medicine, Committee on Systems Approaches to Improve Patient Care by Supporting Clinician Well-Being. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. Washington, DC: National Academies Press (US); 2019.
76. Honavar S. Electronic medical records - the good, the bad and the ugly. Indian J Ophthalmol 2020 Mar;68(3):417-418 [FREE Full text] [doi: 10.4103/ijo.IJO_278_20] [Medline: 32056991]
77. McGreevey J, 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]
78. Kang C, Sarkar IN. Interventions to reduce electronic health record-related burnout: a systematic review. Appl Clin Inform 2024 Jan;15(1):10-25 [FREE Full text] [doi: 10.1055/a-2203-3787] [Medline: 37923381]
79. Turbow S, Hollberg JR, Ali MK. Electronic health record interoperability: how did we get here and how do we move forward? JAMA Health Forum 2021 Mar 01;2(3):e210253 [FREE Full text] [doi: 10.1001/jamahealthforum.2021.0253] [Medline: 36218452]
80. U.S. Department of Health and Human Services. Strategy on reducing burden relating to the use of health IT and EHRsng Burden Relating to the Use of Health IT and EHRs. The Office of the National Coordinator for Health Information Technology. 2020. URL: https://tinyurl.com/4z9fv83y [accessed 2024-04-05]
81. Shah T, Kitts AB, Gold JA, Horvath K, Ommaya A, Frank O, et al. Electronic health record optimization and clinician well-being: a potential roadmap toward action. NAM Perspect 2020 Aug 3;2020:10.31478/202008a [FREE Full text] [doi: 10.31478/202008a] [Medline: 35291737]
82. Gandhi TK, Classen D, Sinsky CA, Rhew DC, Vande Garde N, Roberts A, et al. How can artificial intelligence decrease cognitive and work burden for front line practitioners? JAMIA Open 2023 Oct;6(3):ooad079 [FREE Full text] [doi: 10.1093/jamiaopen/ooad079] [Medline: 37655124]
83. Guo U, Chen L, Mehta PH. Electronic health record innovations: helping physicians - one less click at a time. Health Inf Manag 2017 Sep;46(3):140-144 [FREE Full text] [doi: 10.1177/1833358316689481] [Medline: 28671038]
84. Cohen C, Pignata S, Bezak E, Tie M, Childs J. Workplace interventions to improve well-being and reduce burnout for nurses, physicians and allied healthcare professionals: a systematic review. BMJ Open 2023 Jun 29;13(6):e071203 [FREE Full text] [doi: 10.1136/bmjopen-2022-071203] [Medline: 37385740]
85. Examining clinician burnout in healthcare organizations – why it's also an IT concern. Wolters Kluwer. 2023 Apr 12. URL: https://tinyurl.com/3xa4ynef [accessed 2024-04-05]

Abbreviations

EHR: electronic health record
HIT: health information technology
NASA-TLX: NASA Task-Load Index

Edited by C Lovis; submitted 14.12.23; peer-reviewed by R Schreiber, L Ozeran; comments to author 02.01.24; revised version received 15.02.24; accepted 11.03.24; published 12.04.24.

Please cite as:
Asgari E, Kaur J, Nuredini G, Balloch J, Taylor AM, Sebire N, Robinson R, Peters C, Sridharan S, Pimenta D
Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review
JMIR Med Inform 2024;12:e55499
URL: https://medinform.jmir.org/2024/1/e55499
doi:10.2196/55499
PMID:38607672

Event Analysis for Automated Estimation of Absent and Persistent Medication Alerts: Novel Methodology

Janina A Bittmann[1,*], Dr sc hum; Camilo Scherkl[2,*]; Andreas D Meid[2], PD Dr sc hum; Walter E Haefeli[1], Prof Dr med; Hanna M Seidling[1], Prof Dr sc hum

[1]

[2]

[*]these authors contributed equally

Corresponding Author:
Hanna M Seidling, Prof Dr sc hum

Abstract

Background: Event analysis is a promising approach to estimate the acceptance of medication alerts issued by computerized physician order entry (CPOE) systems with an integrated clinical decision support system (CDSS), particularly when alerts cannot be interactively confirmed in the CPOE-CDSS due to its system architecture. Medication documentation is then reviewed for documented evidence of alert acceptance, which can be a time-consuming process, especially when performed manually.

Objective: We present a new automated event analysis approach, which was applied to a large data set generated in a CPOE-CDSS with passive, noninterruptive alerts.

Methods: Medication and alert data generated over 3.5 months within the CPOE-CDSS at Heidelberg University Hospital were divided into 24-hour time intervals in which the alert display was correlated with associated prescription changes. Alerts were considered "persistent" if they were displayed in every consecutive 24-hour time interval due to a respective active prescription until patient discharge and were considered "absent" if they were no longer displayed during continuous prescriptions in the subsequent interval.

Results: Overall, 1670 patient cases with 11,428 alerts were analyzed. Alerts were displayed for a median of 3 (IQR 1-7) consecutive 24-hour time intervals, with the shortest alerts displayed for drug-allergy interactions and the longest alerts displayed for potentially inappropriate medication for the elderly (PIM). Among the total 11,428 alerts, 56.1% (n=6413) became absent, most commonly among alerts for drug-drug interactions (1915/2366, 80.9%) and least commonly among PIM alerts (199/499, 39.9%).

Conclusions: This new approach to estimate alert acceptance based on event analysis can be flexibly adapted to the automated evaluation of passive, noninterruptive alerts. This enables large data sets of longitudinal patient cases to be processed, allows for the derivation of the ratios of persistent and absent alerts, and facilitates the comparison and prospective monitoring of these alerts.

(JMIR Med Inform 2024;12:e54428) doi:10.2196/54428

KEYWORDS

clinical decision support system; CDSS; medication alert system; alerting; alert acceptance; event analysis

Introduction

Computerized physician order entry (CPOE) systems with integrated clinical decision support systems (CDSS) can reduce medication errors by highlighting critical medication constellations [1]. To realize their full potential, medication alerts must be recognized and followed by users. Hence, measuring "alert acceptance" is a key prerequisite for evaluating the effectiveness of a CDSS.

In principle, two methods can estimate alert acceptance: (1) in-dialog analysis where users interactively click to accept or override displayed alerts; and (2) event analysis where the medication chart and associated documentation are reviewed for evidence of alert acceptance through further actions ("events") responsive to the alert (eg, discontinued medication orders), which often requires extensive manual screening [2]. Most studies addressing alert acceptance used in-dialog analyses because the display of alerts, especially in English-speaking countries, is part of the technical architecture of the CPOE-CDSS [3]. There is limited evidence on how to perform event analyses because it is uncertain whether the prescribing behavior is influenced by alerts or other clinical therapeutic circumstances (eg, scheduled end of treatment) [2]. Moreover, the manual screening of the medication documentation is a time-consuming process [4], especially when administrative

processes such as changing wards and the simultaneous transfer of physicians' responsibility for the medication are considered in the alert presentation.

As CDSS installations presenting passive, noninterruptive alerts become increasingly popular in European countries [5,6], the need for developing and validating techniques for automatic event analyses is increasing. This is particularly important when considering all alerts throughout the inpatient stay.

We present a new approach to perform an automated event analysis, which was applied to a large data set of medication alerts.

Methods

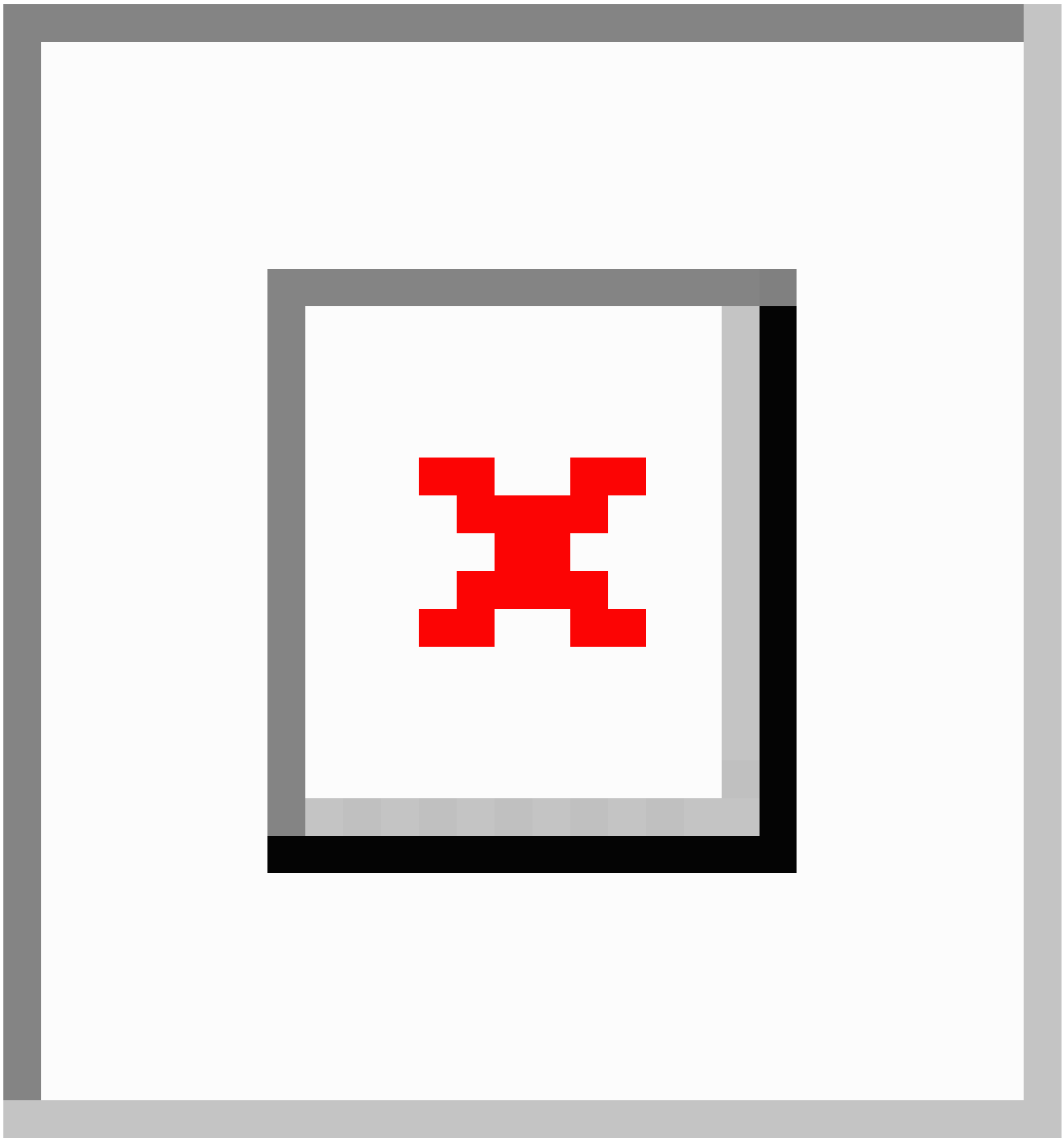
Ethical Considerations

Study approval was granted by the responsible Ethics Committee of the Medical Faculty of Heidelberg University (S-467/2020) and by the local data protection officer for the data protection concept. Human subjects were not directly involved; all data were pseudonymized and could neither be linked to individual patients nor to prescribers.

Setting

We analyzed the prescription and alert data issued over 3.5 months during routine care at Heidelberg University Hospital (a 2500-bed tertiary care hospital) within the local CPOE (*i.s.h.med Smart Medication*, Oracle Cerner, North Kansas City, USA) with an integrated CDSS (AiD*Klinik*, Dosing GmbH, Heidelberg, Germany). To view the presented passive and noninterruptive alerts, users actively navigate from their prescription screen to a separate window that opens upon request. In this window, all alerts are displayed in a single table sorted by severity and presented with a brief summary (Figure 1). Users are required to click on each alert to access more detailed information. The system does not recognize whether an alert has been viewed. Additionally, users are not obliged to directly flag alerts as accepted or overwritten. Therefore, these data are not available in our CPOE-CDSS. Implemented alert types comprised checking for drug-drug interactions (DDIs), drug-allergy interactions (DAIs), duplicate prescriptions (DPs), advanced dosing recommendations for potentially inappropriate medication for the elderly (aged ≥65 years, PIM), or prescriptions exceeding the maximum recommended daily dose (PE-MDDs).

Figure 1. Schematic display of an exemplary alert window listing all alerts for a patient in a table. Each alert is presented in a separate line, sorted by severity, with the most severe alerts listed first. The first column displays the alert type in a color-coded scheme (black=contraindicated, red=severe, orange=moderate), followed by an explanation for the severity, a brief description of the problem, and the name of the causative drug.



Data Collection

The relevant parameters extracted from the CPOE-CDSS were information on prescriptions, issued alerts, administrative patient data, and setting data. Prescription schedules with regimen changes were documented as separate entries so that prescriptions potentially resulting from previous prescriptions (eg, because of dose reduction or conversion of fixed to as-needed prescriptions) could be linked retroactively. Follow-up prescriptions were defined as prescriptions of the same drug and administration route when the previous prescription ended and the subsequent one started within 10 minutes.

Alert Management

In this CDSS, prescriber review of alerts may result in alerts disappearing due to prescription changes and adaptations or in alerts being continuously displayed for unchanged prescriptions.

In this methodology, alerts are defined as “absent” when they disappear during continuous prescriptions for which underlying risk constellations no longer exist (eg, dose reduction of an overdosed prescription but the prescription itself remains continuous). In contrast, alerts consistently displayed until patient transfer, discharge, or end of the prescription are categorized as “persistent” (eg, the prescription remains valid

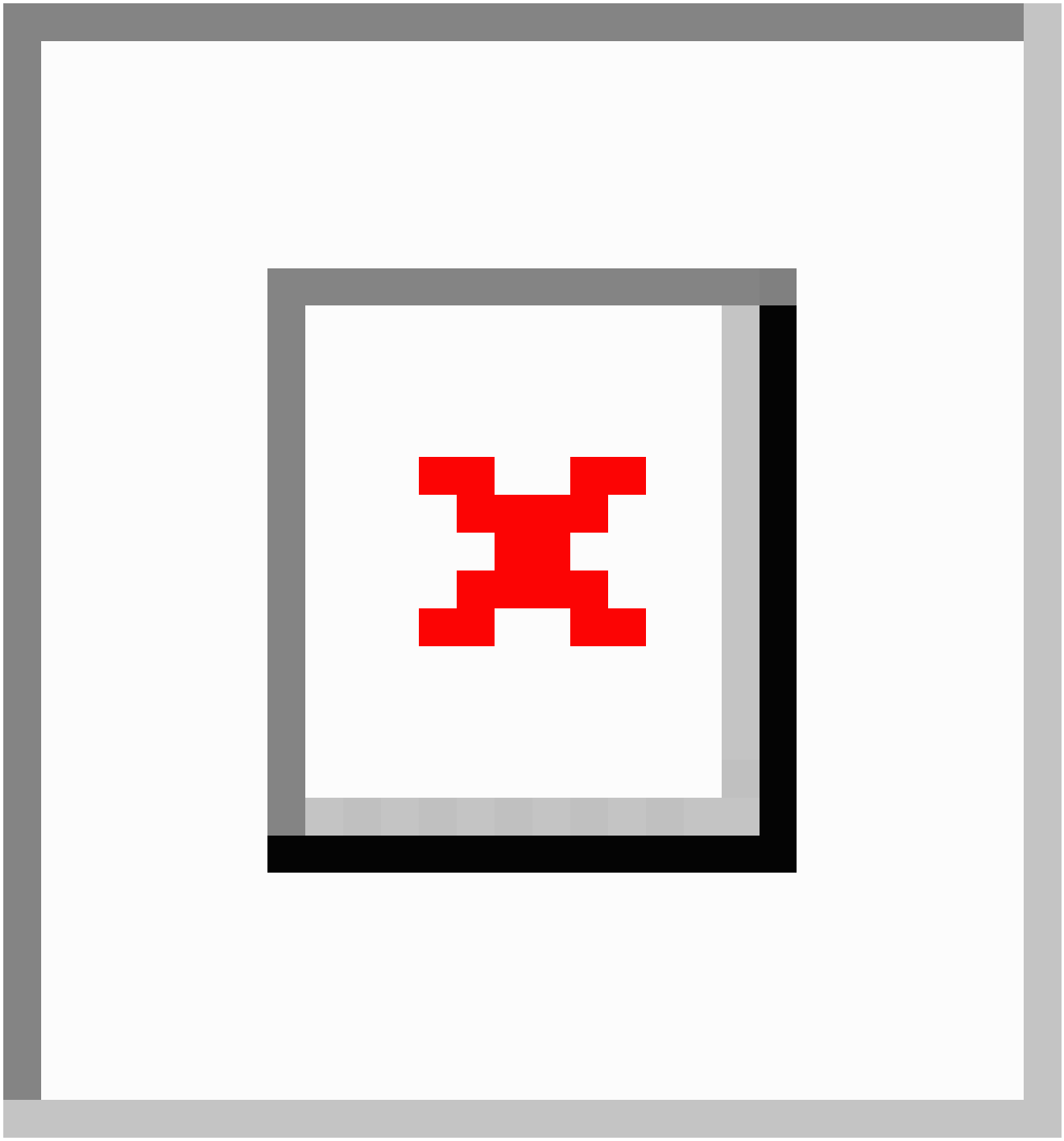
even though the prescribed active ingredient is alerted by a DAI).

Data Analysis

To automatically identify absent alerts, the medication and corresponding alert data were divided into 24-hour time intervals. Since reproducible time stamps were lacking in the database, selection of this conservative interval allowed the retroactive linkage of alert display and associated changes in the corresponding prescription. Therefore, two consecutive time intervals were compared according to whether or not the alerts were continuously displayed. Alerts were excluded if (1) they were first displayed on the discharge day (Figure 2, Alert 4); (2) required user interaction (eg, answering questions to decide whether conditions for alerts are present); or (3) were triggered by one-time prescriptions, in which case the alert response cannot be assessed the next day (Supplementary Methods section A and Figure S1 in Multimedia Appendix 1). The display duration of alerts (DDoA) was calculated from the time interval between the first and last alert display (Figure 2). Further details on the development of the 24-hour time intervals can be found in Supplementary Methods section B in Multimedia Appendix 1; examples of data analysis of different alert types are provided in Tables S1-S3 and Figure S2 of Multimedia Appendix 1.

Based on the exploratory rates of absent alerts, basic descriptive statistics were applied. The χ^2 test was performed to evaluate whether absent alerts differed stratified by alert types considering a two-tailed *P* value <.05 as significant (IBM SPSS Statistics version 25, Ehningen, Germany). The R packages *gpmodels* [7], *survival*, and *ggplot2* (version 4.1.2, R Foundation for Statistical Computing, Vienna, Austria) were used for data analysis and visualization.

Figure 2. Proposed methodology for identification of absent alerts, exemplified for a 5-day inpatient stay. Each midnight (dotted lines), all alerts displayed within the last 24-hour time interval were identified. Alert 1 is displayed between day 1 (admission) and day 4; the display duration of alerts (DDoA) is 4 days. Alert 2 is displayed from day 2 until discharge; the DDoA is 4 days. Alert 3 is displayed only on day 3; the DDoA is 1 day. Alert 4 is displayed for the first time on the day of discharge and remained until discharge; the DDoA is 1 day. Alert 4 was excluded from the analysis because, due to the discharge, there is no subsequent (sixth) 24-hour time interval with which the fifth interval could have been compared to evaluate the alert display. Each alert could be identified because of a unique alert ID code. Using this identification concept, alert IDs detected within a 24-hour time interval could be compared to alerts detected in the previous 24-hour time interval. Therefore, it was possible to automatically classify which alerts were (1) newly displayed (no matching ID in the previous interval: Alert 1, Day 1), (2) displayed for more than 24 hours (matching ID in consecutive intervals; Alert 1, Days 2-4), or (3) absent (no matching ID in the current 24-hour time interval: Alert 1, Day 4).



Results

Alert Display and Composition

We considered the data of 1670 patient cases (Figure S3 in Multimedia Appendix 1) with a median hospital stay of 7 days (IQR 4-13). During this time, 13,979 alerts were displayed. Because 2284 alerts (16.3%) were triggered by one-time prescriptions and 267 alerts (1.9%) were first displayed on the discharge day, the remaining 11,428 alerts (81.8%) formed the basis for analysis. The alert types triggering the alerts are shown in Table 1.

The median DDoA was 3 days (IQR 1-7) and varied by alert type, with alerts for DAIs showing the shortest DDoA (Table 1).

Table . Alert types triggering the alerts, corresponding rates of absence, and display duration of the alerts.

Alert type	Triggered alerts, n (%)[a]	Absent alerts, n (%)[b]	Display duration of alerts (days), median (IQR; range)
Alerts for duplicate prescriptions	7643 (66.9)	3674 (48.1)	3 (1-8; 1-31)
Alerts for drug-drug interactions	2366 (20.7)	1915 (80.9)	2 (1-5; 1-31)
Alerts for drug-allergy interactions	517 (4.5)	416 (80.5)	1 (1-2; 1-24)
Alerts for potentially inappropriate medication for the elderly	499 (4.4)	199 (39.9)	4 (2-8; 1-31)
Alerts for prescriptions exceeding the maximum recommended daily dose	403 (3.5)	209 (51.9)	3 (1-6; 1-30)
Total number of alerts	11,428	6413	3 (1-7; 1-31)

[a]Percentages are based on the total number of analyzed alerts (N=11,428).

[b]Percentages are based on the number of analyzed alerts for each alert type.

Absent and Persistent Alerts

From all 11,428 analyzed alerts, 43.9% (n=5015) persisted and 56.1% (n=6413) were absent, with alerts for DDIs showing the highest rate of absence (80.9%) and PIMs the lowest (39.9%) (Table 1).

The proportions of absent alerts differed significantly between the individual alert types (P_{χ^2}<.001), except for DDI alerts compared to DAI alerts (P_{χ^2}=.80) and for DP alerts compared to alerts for PE-MDDs (P_{χ^2}=.14). The proportion of absent alerts in relation to the DDoA was the highest for DAI alerts and the lowest for alerts for PIMs in the first 24 hours after admission (Figures S4-S5 in Multimedia Appendix 1).

Discussion

Principal Findings

A new methodological approach for routine care data was applied performing an automated event analysis that is transferable to other CPOE-CDSS with passive, noninterruptive alerts. In previous studies using event analyses, alert acceptance rates were identified at the drug administration level [8], prescription level [9], or at both levels [10]. There is general consensus that alert acceptance rates vary widely depending on the measuring method and study setting, resulting in different and incomparable rates [11]. Since in-dialog analysis is often not possible in a European CPOE-CDSS, this new methodology adapted to the technical structures of such a CPOE-CDSS is needed.

A key strength of the proposed method is that it variably adjusts the time intervals and consequently the lookback windows underlying the method's programming. Thus, temporary changes in prescriptions within the determined time interval (here 24 hours) are considered persistent alerts; however, this CPOE-CDSS interrupts the alert display in certain cases, such as when patients change wards and responsibility for medication is handed over to another physician. This transfer results in automatic prescription pauses that are actively suspended by physicians, technically leading to the redisplay of alerts. Without the definition of this time interval, these pauses would incorrectly increase the overall number of alerts when reappearing and the rate of absent alerts as they disappear for a few hours during valid prescriptions. Hence, this method considers administrative processes of the daily clinical routine and guarantees that only alerts of real prescription changes are evaluated. For retrospectively matching the time-dependent correlations of the alerts over time and in the clinical routine, it is essential to consider alerts throughout the inpatient stay and our proposed method meets this need.

However, according to the technical architecture of this CPOE-CDSS, there is no obvious link between reviewing alerts and possible resulting changes in prescription data. Therefore, various assumptions had to be made for this data evaluation. Alerts were categorized as either persistent or absent based on the assumptions that alerts were regularly checked and that alerts disappeared because underlying risk constellations no longer existed due to previously displayed alerts. This general assumption may overestimate the rate of actual alert acceptance, as a prescribed medication could be switched based on patient conditions (eg, adverse events, intolerance) or treatment schedules. As it remains unclear whether the change in drug prescriptions was caused by the alert display or due to other variables and because no control group was available due to the retrospective design, caution is required when interpreting absolute numbers and comparing the proportion of absent alerts to previously published acceptance rates. In the future, this retrospective method will need to be prospectively evaluated including validity measurements by comparing the results of this automated approach with those derived from manual screening. Another limitation is that this study was conducted in a single center with a CPOE-CDSS that is highly specific and strongly adapted to workflows and care processes at our institution. This analysis only considered alerts at the prescribing level and did not measure whether the respective drugs were indeed administered. For instance, many of the alerts for DPs were triggered by drugs that were prescribed as as-needed prescriptions. Hence, these DPs tended to indicate a variety of treatment options rather than actually being administered together. This might have contributed to the reduced occurrence of the absence of alerts for DPs on a prescribing level compared to other alerts. However, in our CPOE-CDSS, it is unalterably

stipulated that medication alerts are implemented in a passive and noninterruptive way. While it may be challenging to transfer this complex method to systems with differing data infrastructures, to our knowledge, this is the first automated method for processing persistent and absent medication alerts in a system with passive, noninterruptive alerts. Additionally, since this method was programmed in a modular way, it seems feasible to transfer and adapt it to other settings.

Conclusions

A methodology was applied to an automatic event analysis in a CPOE-CDSS with passive, noninterruptive alerting. This enables the processing of large data sets of longitudinal periods of inpatient stays and can be used to automatically derive the percentage of absent alerts. Once implemented, this analysis can be repeated at any time and one could even imagine that real-time monitoring of persistent alerts in daily clinical routines could be set up using these data for future optimization of the CPOE-CDSS.

Acknowledgments

We thank Sonja Baumann, Silvia Kugler, Michael Metzner, Larissa Schiller, and Andreas Wirtherle for initial data extraction, preparation, and maintenance.

Authors' Contributions

JAB planned the study, was involved in the development of the method and data evaluation, and wrote the manuscript. CS was involved in the development of the method and data evaluation and wrote the manuscript. ADM was involved in the development of the method and data evaluation, wrote parts of the manuscript and critically revised it. WEH wrote parts of the manuscript and critically revised it. HMS planned the study, was involved in the development of the method and data evaluation, and wrote and critically revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Description of one-time prescription and illustration of its impact on the alert display (Figure S1). Description of the development of 24-hour time intervals, with exemplary data sets for fixed variables (Table S1), time-dependent variables (Table S2), and an exemplary time frame of processed longitudinal alert data (Table S3). Data analysis examples for different alert types (Figure S2). CONSORT (Consolidated Standards of Reporting Trials) diagram for included patient cases (Figure S3). Proportions of absent (Figure S4) and persistent (Figure S5) alerts stratified by the alert type.
[DOCX File, 353 KB - medinform_v12i1e54428_app1.docx]

References

1. Wolfstadt JI, Gurwitz JH, Field TS, et al. The effect of computerized physician order entry with clinical decision support on the rates of adverse drug events: a systematic review. J Gen Intern Med 2008 Apr;23(4):451-458. [doi: 10.1007/s11606-008-0504-5] [Medline: 18373144]
2. Hussain MI, Reynolds TL, Zheng K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. J Am Med Inform Assoc 2019 Oct 1;26(10):1141-1149. [doi: 10.1093/jamia/ocz095] [Medline: 31206159]
3. Duke JD, Li X, Dexter P. Adherence to drug-drug interaction alerts in high-risk patients: a trial of context-enhanced alerting. J Am Med Inform Assoc 2013 May 1;20(3):494-498. [doi: 10.1136/amiajnl-2012-001073] [Medline: 23161895]
4. Sethuraman U, Kannikeswaran N, Murray KP, Zidan MA, Chamberlain JM. Prescription errors before and after introduction of electronic medication alert system in a pediatric emergency department. Acad Emerg Med 2015 Jun;22(6):714-719. [doi: 10.1111/acem.12678] [Medline: 25998704]
5. Carli-Ghabarou D, Seidling HM, Bonnabry P, Lovis C. A survey-based inventory of clinical decision support systems in computerised provider order entry in Swiss hospitals. Swiss Med Wkly 2013;143:w13894. [doi: 10.4414/smw.2013.13894] [Medline: 24338034]
6. Ploegmakers KJ, Medlock S, Linn AJ, et al. Barriers and facilitators in using a clinical decision support system for fall risk management for older people: a European survey. Eur Geriatr Med 2022 Apr;13(2):395-405. [doi: 10.1007/s41999-021-00599-w] [Medline: 35032323]
7. R package gpmodels. GitHub. URL: https://github.com/ML4LHS/gpmodels [accessed 2023-11-16]
8. Muylle KM, Gentens K, Dupont AG, Cornu P. Evaluation of an optimized context-aware clinical decision support system for drug-drug interaction screening. Int J Med Inform 2021 Apr;148:104393. [doi: 10.1016/j.ijmedinf.2021.104393] [Medline: 33486355]

9. Slight SP, Beeler PE, Seger DL, et al. A cross-sectional observational study of high override rates of drug allergy alerts in inpatient and outpatient settings, and opportunities for improvement. BMJ Qual Saf 2017 Mar;26(3):217-225. [doi: 10.1136/bmjqs-2015-004851] [Medline: 26993641]
10. Muylle KM, Gentens K, Dupont AG, Cornu P. Evaluation of context-specific alerts for potassium-increasing drug-drug interactions: a pre-post study. Int J Med Inform 2020 Jan;133:104013. [doi: 10.1016/j.ijmedinf.2019.104013] [Medline: 31698230]
11. Kannry J. Alert acceptance: are all acceptance rates the same? J Am Med Inform Assoc 2023 Sep 25;30(10):1754. [doi: 10.1093/jamia/ocad151] [Medline: 37535817]

Abbreviations

CDSS: clinical decision support system
CPOE: computerized physician order entry
DAI: drug-allergy interaction
DDI: drug-drug interaction
DDoA: display duration of alert
DP: duplicate prescription
PE-MDD: prescription exceeding the maximum recommended daily dose
PIM: potentially inappropriate medication for the elderly

Edited by C Lovis; submitted 17.11.23; peer-reviewed by A Simona, D Malone; revised version received 19.03.24; accepted 07.04.24; published 04.06.24.

Please cite as:
Bittmann JA, Scherkl C, Meid AD, Haefeli WE, Seidling HM
Event Analysis for Automated Estimation of Absent and Persistent Medication Alerts: Novel Methodology
JMIR Med Inform 2024;12:e54428
URL: https://medinform.jmir.org/2024/1/e54428
doi:10.2196/54428

Original Paper

Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation Use Case: Implementation and Evaluation Study

Boshu Ru[1], PhD; Arthur Sillah[1], MPH, PhD; Kaushal Desai[1], MS, PhD; Sheenu Chandwani[1], MPH, PhD; Lixia Yao[1], PhD; Smita Kothari[1], MBA, PhD

Center for Observational and Real-world Evidence (CORE), Merck & Co, Inc, West Point, PA, United States

Corresponding Author:
Boshu Ru, PhD
Center for Observational and Real-world Evidence (CORE)
Merck & Co, Inc
770 Sumneytown Pike
WP37A
West Point, PA, 19486
United States
Phone: 1 215 652 4301
Email: boshu.ru@merck.com

Abstract

Background: The importance of real-world evidence is widely recognized in observational oncology studies. However, the lack of interoperable data quality standards in the fragmented health information technology landscape represents an important challenge. Therefore, adopting validated systematic methods for evaluating data quality is important for oncology outcomes research leveraging real-world data (RWD).

Objective: This study aims to implement real-world time to treatment discontinuation (rwTTD) for a systemic anticancer therapy (SACT) as a new use case for the Use Case Specific Relevance and Quality Assessment, a framework linking data quality and relevance in fit-for-purpose RWD assessment.

Methods: To define the rwTTD use case, we mapped the operational definition of rwTTD to RWD elements commonly available from oncology electronic health record–derived data sets. We identified 20 tasks to check the completeness and plausibility of data elements concerning SACT use, line of therapy (LOT), death date, and length of follow-up. Using descriptive statistics, we illustrated how to implement the Use Case Specific Relevance and Quality Assessment on 2 oncology databases (*Data sets A and B*) to estimate the rwTTD of an SACT drug (*target SACT*) for patients with advanced head and neck cancer diagnosed on or after January 1, 2015.

Results: A total of 1200 (24.96%) of 4808 patients in Data set A and 237 (5.92%) of 4003 patients in Data set B received the target SACT, suggesting better relevance of the former in estimating the rwTTD of the target SACT. The 2 data sets differed with regard to the terminology used for SACT drugs, LOT format, and target SACT LOT distribution over time. Data set B appeared to have less complete SACT records, longer lags in incorporating the latest data, and incomplete mortality data, suggesting a lack of fitness for estimating rwTTD.

Conclusions: The fit-for-purpose data quality assessment demonstrated substantial variability in the quality of the 2 real-world data sets. The data quality specifications applied for rwTTD estimation can be expanded to support a broad spectrum of oncology use cases.

(JMIR Med Inform 2024;12:e47744) doi:10.2196/47744

KEYWORDS

data quality assessment; real-world data; real-world time to treatment discontinuation; systemic anticancer therapy; Use Case Specific Relevance and Quality Assessment; UReQA framework

Introduction

Background

The importance of real-world evidence drawn from real-world data (RWD) is widely recognized in oncology research [1-5]. Over the past decade, federal legislation and incentives promoting the secondary use of RWD in the United States [6-8], coupled with advances in health information technology, have resulted in an explosion of RWD sources and a complex RWD ecosystem [1]. However, this rich data landscape can also pose challenges in identifying fit-for-purpose RWD to meet biopharma research needs.

Two key obstacles to identifying high-quality data are the fragmentation of RWD sources and the lack of interoperable data quality standards. These obstacles are particularly pertinent in the United States, where progress is slow in reaching full interoperability of data sourced from thousands of providers who customized their electronic health record (EHR) systems from solutions provided by >40 different EHR software vendors [9]. Therefore, adopting validated systematic methods for evaluating data quality is important for research leveraging RWD [10-12].

In 2016, an expert panel proposed the concepts of *conformance*, *completeness*, and *plausibility* as 3 categories (with subcategories) to describe the intrinsic data quality of EHR databases and to serve as a framework for assessing data quality that could then be verified (with organizational data) or validated using an accepted gold standard [13]. Several working groups and authors have applied these terms or proposed others for defining research data quality [14-16], and multiple initiatives in the United States, both public and private, have developed frameworks and tools to evaluate and improve the quality of EHR data sets [17-21] and to implement model-driven, quantitative approaches to address RWD completeness and plausibility issues [22-25]. However, there is no single RWD source that can fit the needs of all studies, and the selection of RWD to support an individual use case must also consider data relevance and measurement thresholds in addition to data quality.

Objective

In a previous study, we introduced the Use Case Specific Relevance and Quality Assessment (UReQA) framework, an RWD quality framework that combines both the data quality and the relevance aspects of assessing RWD, with the goal of developing data quality assessment specifications tailored to use cases [3]. In this study, we aimed to implement this framework in the use case for estimating real-world time to treatment discontinuation (rwTTD) in oncology. Our work had two main components: (1) to design comprehensive data quality assessment checks for estimating rwTTD for a systemic anticancer therapy (SACT) and (2) to illustrate how these quality checks can be used to evaluate EHR-derived RWD products.

We selected rwTTD as the first use case to implement the UReQA framework because of its high utility as a pragmatic real-world effectiveness end point for continuously administered SACTs (such as immunotherapies) and its known correlation with overall survival [26-28]. Moreover, the estimation of rwTTD requires information on medication use patterns, mortality, and follow-up. These data elements are foundational to outcomes research. Therefore, implementation of the rwTTD use case can be expanded to other use cases in or beyond oncology, as well as different data sources, such as claims databases.

Methods

Ethical Considerations

This study used 2 commercially licensed deidentified structured secondary data sources accessible to the study team. It was exempted from institutional review board review because of the following: (1) each data source contains a significant level of protection against the release of personal information to outside entities and (2) the use of such databases presents the lowest risk to potential subjects because the analysis involves only anonymous data; hence, conducting the study will not place the subjects at risk.

Study Overview

This study comprised four main steps: (1) conceptual definition of the rwTTD use case; (2) mapping of the rwTTD use case definition to RWD elements (operational definition); (3) identifying data quality checks for the required data elements to determine rwTTD for an SACT, designated the “target SACT”; and (4) implementing the UReQA framework [3] in assessing the RWD fitness for estimating rwTTD. The data quality assessment was undertaken on 2 US EHR-based oncology databases for estimating rwTTD for a target SACT, an immunotherapy drug that is administered intravenously in advanced-stage head and neck cancer (HNC). The targeted SACT received approval in 2016 for the treatment of previously treated advanced HNC and in 2019 for its use as a first-line therapy in advanced HNC. The focus of this study is on designing data quality assessment methods that are tailored for specific use cases, rather than calculating rwTTD for a particular medication. Therefore, we mask the name of the actual drug product.

Step 1: Conceptual Definition of the rwTTD Use Case

The end point, rwTTD, is defined as the length of time from initiation to discontinuation of a medication (*[date of last recorded dose – date of first recorded dose] + 1 d*), with discontinuation defined as the date of the last dose if a patient died during therapy or initiated a new treatment or if there is a gap of ≥120 days between the last recorded dose and last recorded activity in a data set. Patients who do not meet the discontinuation criteria are censored at the last medication use [26-28].

Step 2: Mapping of the rwTTD Use Case Definition to RWD Elements

Owing to the variations in data element definition and data structures between real-world EHR databases, we need to operationalize the concept of rwTTD by deconstructing its definition and mapping it to four sets of required data elements that are commonly available from oncology EHR–derived data

sets: (1) SACT, (2) line of therapy (LOT) specifying the regimen names and sequence of current treatment in the treatment plan [29,30], (3) mortality status, and (4) follow-up time, as summarized in Table 1. Although SACT, mortality status, and follow-up time are often recorded directly as procedure, prescription, and administrative events in raw EHR databases, the LOT was often derived from raw EHR by the algorithm.

Table 1. Required data elements for determining real-world time to treatment discontinuation (rwTTD) for a systemic anticancer therapy (SACT) drug in a specific line of therapy (LOT).

Operational steps to ascertain rwTTD and type of data category	Commonly used data elements
Identify records of the drug of interest	
SACT drug	Drug_name, NDC[a], HCPCS[b] code, RxNorm code
SACT administration	Drug administration date[c]
SACT order	Drug order date[d]
Identify discontinuation date from subsequent LOT start date	
LOT	LOT name[e]
LOT	LOT number
LOT	LOT start date
LOT	LOT end date
If no subsequent LOT, identify discontinuation date from patient death record during treatment	
Mortality status	Vital status or date of death
If no date of death, identify discontinuation date by last follow-up date subheading	
Last follow-up	Date of last follow-up[f]

[a]NDC: National Drug Code.

[b]HCPCS: Healthcare Common Procedure Coding System.

[c]The drug administration date is defined as the date of receiving medication at a health care facility as a medical service, often applicable to an intravenous drug.

[d]The drug order date is defined as the order date for drugs used at home, often applicable to an oral drug.

[e]The LOT name is determined by the combination of SACT drugs administered or ordered from the LOT start to end dates.

[f]The date of last follow-up is defined as the last documented clinic visit or procedure in the electronic health record.

We defined SACT as any systemic anticancer medication received by the patient, documented as given either by a health care provider at the site of care (eg, by infusion), with the date defined as the "administration" date, or as a prescription to take or apply at home, with the date defined as the "drug order" date. The number of refills (or alternative data elements such as days of supply or expected medication end date) was used to determine the last use of oral drugs (Table 1).

LOT was defined as the sequence of the SACT regimens prescribed for an individual patient, as previously described in detail [29,30]. In brief, the first LOT (line 1 [1L]) begins with the first SACT initiated after a study index date (often the advanced or metastatic cancer diagnosis date), and any other drug introduced within the next 28 days is considered part of that LOT [29]. We defined the start of a new LOT when a new SACT not belonging to the prior LOT was introduced or if a new SACT was initiated after a ≥120-day gap in therapy.

Because the target SACT was administered intravenously, we omitted 2 tasks applicable only to oral target SACTs: the check of patient numbers with target drug order date after the index date (Multimedia Appendix 1, task 6 [13]) and the check for distribution of gaps between drug order dates (Multimedia Appendix 1, task 9).

The patient mortality status was determined based on the recorded dates of death. For patients who were still alive at data cutoff, the date of the last follow-up was defined as the last documented clinical activity date in the EHR (Table 1).

Step 3: Identifying Data Quality Checks for Required Data Elements

For each of the required data elements, we identified corresponding verification checks to assess data quality at both the variable level and the cohort level. A total of 20 data quality checks (tasks) were identified and categorized into the quality dimensions of conformance, completeness, and plausibility, as per the harmonized data quality assessment terms and framework developed by Kahn et al [13] (Multimedia Appendix 1). Our goal in creating these tasks was to develop a comprehensive toolbox for assessing data quality for the rwTTD use case. However, when adapting them to a specific RWD database and a SACT drug of interest, not every task and check would be necessary. For example, the checks for LOT, mortality, and follow-up are not needed if a data set already provides the reason for discontinuation and censored status for each drug exposure. In addition, tasks 3-5 were applicable to cancer therapies received in hospitals or clinics as intravenous or infusion procedures, whereas tasks 4-9 were dedicated to oral

cancer therapies that were mostly self-administrated at home. As tracking the actual time patients took oral therapies was infeasible, researchers examined days supply and refill records to estimate the drug exposure period. Therefore, when investigating the rwTTD of an oral SACT drug, it is necessary to check the completeness of these oral therapy–specific data elements (task 7).

Step 4: Implementing the rwTTD Use Case for Assessing 2 RWD Sets

Data Set Preassessment

We followed the preassessment step in UReQA [3] to identify 2 anonymized, commercially available US real-world oncology databases, designated as *Data set A* and *Data setB* in this report, which included patients with advanced (metastatic or unresectable, recurrent) HNC. Both databases contained data elements sourced from structured and unstructured information captured within health care providers' EHR systems as part of routine cancer care.

Cohort Selection and Patient Characteristics

Data set A was commercialized and included patients with advanced HNC, whereas Data set B included patients with all stages of HNC. To align the 2 patient populations as having advanced HNC, we restricted Data set B to the subset of patients with HNC and a record of the American Joint Committee on Cancer stage IV and *International Classification of Diseases* (ICD), *revision 9 or 10* (ICD-9 or ICD-10) code for metastatic tumor (ICD-9 codes 196.x, 197.x, and 198.x and ICD-10 codes C76.x, C77.x, and C78.x). The distributions of the patient characteristics were then tabulated for the 2 data sets.

Data Elements Harmonization

In Data set A, the names of SACT medications were harmonized from clinic formulary information and medical service records to standard generic drug names in a commercial drug database along with drug category information. In Data set B, all medication records in the raw EHR data were harmonized into the RxNorm code [31]; however, drug category information was not available. To harmonize all SACT medication in Data set B, we retrieved the RxNorm codes for generic names of all SACT medications using the RxNav software developed by and available from the US National Library of Medicine [32].

The LOT information was previously derived by both data providers but was presented differently in the 2 data sets. In Data set A, the LOT table provided the LOT number, LOT regimen name, LOT start date, and LOT end date, with a flag indicative of maintenance therapy, as appropriate. Instead, Data set B included only the LOT number and LOT start date. Therefore, to evaluate the LOT information in Data set B, we indirectly deduced the end date of each LOT as the date before the start of the next LOT or as the data cutoff date for the last LOT in the data set. Then, all individual SACT medications administered or ordered between the LOT start and end dates were combined to serve as the LOT regimen name. This approach was a necessary but imperfect solution because the LOT end date and the LOT regimen name should ideally be generated using a more rigorous algorithm [29,30].

The date of death was provided at the month and day levels in Data set A, whereas in Data set B, the death date was aggregated by year. Given the relatively short length of survival of many patients with advanced HNC [33-36], the allocation of death dates by year was not sufficiently granular for accurate rwTTD calculation; better precision (ie, month of death) would be needed for accurate rwTTD calculation. Consequently, quality assessment tasks related to mortality variables were omitted (task 17) for Data set B.

Reporting the Verification Results

Descriptive statistics were used to summarize the results of implementing rwTTD data quality checks on Data sets A and B. We used frequencies to summarize categorical variables and mean (SD) and median (IQR or range) to summarize continuous variables. The study index date was the date of first advanced HNC diagnosis, and the cutoff date was November 25, 2019.

All analyses were conducted using SAS Studio release 3.8 (Basic Edition; SAS Institute, Inc).

Results

Patient Characteristics

Data set A included 7366 patients with advanced HNC, and we identified 11,386 patients in Data set B with advanced HNC. The median patient age at the first advanced HNC diagnosis was 65 (IQR 58-72) years in Data set A and 61 (IQR 54-68) years in Data set B, and the percentages of male individuals were 74.16% (5643/7366) and 69.97% (7967/11386), respectively (Table 2), similar to the HNC population data from the United States [33,37].

Table 2. Baseline characteristics of patients with advanced head and neck cancer (HNC) included in 2 data sets under evaluation[a].

Characteristic	Data set A (n=7366)	Data set B (n=11,386)
Sex, n (%)		
Female	1723 (23.4)	3408 (29.9)
Male	5643 (76.6)	7967 (70)
Missing or unknown	0 (0)	11 (0.1)
Age at first advanced HNC diagnosis (y), median (IQR)	65 (58-72)	61 (54-68)
Age at first advanced HNC diagnosis (y), n (%)		
<18	0 (0)	31 (0.27)
18-44	187 (2.53)	688 (6.04)
45-64	3402 (46.19)	5955 (52.3)
65-88	3777 (51.28)	4111 (36.11)
≥89	0 (0)	6 (0.05)
Missing or unknown	0 (0)	595 (5.23)
Race or ethnicity, n (%)		
American Indian or Alaska Native	N/A[b]	40 (0.35)
Asian	103 (1.4)	165 (1.45)
Black or African American	487 (6.61)	1250 (10.98)
Hispanic or Latino	13 (0.18)	0 (0)
Native Hawaiian or other Pacific Islander	N/A	6 (0.05)
White	4939 (67.05)	9239 (81.14)
Missing	650 (8.82)	686 (6.02)
Other race	1174 (15.94)	0 (0)
AJCC[c] stage at first HNC diagnosis, n (%)		
0	2 (0.03)	28 (0.25)
I	419 (5.69)	603 (5.3)
II	505 (6.86)	542 (4.76)
III	929 (12.61)	798 (7.01)
IV	4330 (58.78)	4978 (43.72)
Missing or unknown	1181 (16.03)	4437 (38.97)
Year of first advanced HNC diagnosis, n (%)		
Before 2006	0 (0)	1245 (10.9)
2006-2009	0 (0)	1537 (13.5)
2010-2012	1068 (14.5)	2721 (23.9)
2013-2018	5435 (73.8)	5577 (49.0)
2019 or later	863 (11.7)	306 (2.7)

[a]Percentages may not add up to 100% because of rounding.

[b]N/A: not applicable.

[c]AJCC: American Joint Committee on Cancer.

SACT Data Checks

Overall, 75.91% (5592/7366) and 38.74% (4411/11386) of the patients in Data sets A and B, respectively, had a recorded drug administration or drug order for any SACT (Table 3, task 1). A complete start date (y, mo, and d) was recorded for all SACT administrations and orders in both data sets (Table 3, tasks 4 and 8).

Table 3. Data quality assessment of SACT[a] administration and order records after the advanced HNC[b] diagnosis[c].

SACT data quality checks	Data set A	Data set B
Task 1: patients with any SACT drug administration or order record after the advanced HNC diagnosis date, n (%)[d]	5592 (75.9)	4411 (38.7)
Task 2: SACT drug records with missing drug identity (name and code) information		
Value, n (%)	0 (0)	0 (0)
Normalization of medication name	Normalized generic name	RxNorm ingredient level
Task 3: patients with target SACT administration date after the advanced HNC diagnosis date, 2015 onward, % (n/N)[e]	24.96 (1200/4808)	5.92 (237/4003)
Task 4: SACT drug administration records with complete administration date, n (%)	425,505 (100)	37,662 (100)
Task 5: gap (in d) between the target SACT drug administration dates, median (IQR; range)	21 (21-21; 1-113)	21 (11-21; 1-824)
Task 6: patients with target SACT order date after the advanced HNC diagnosis date	N/A[f,g]	N/A[g]
Task 7 SACT drug order records with complete days supply and refill information, n (%)	1732 (53.4)	N/A[h]
Task 8: SACT drug order records with complete order date, n (%)[i]	3241 (100)	8380 (100)
Task 9: distribution of gaps (in d) between target SACT drug order dates, normalized by days supply, refill, and cancelation record	N/A[g]	N/A[g]

[a]SACT: systemic anticancer therapy.

[b]HNC: head and neck cancer.

[c]Drug *administration* refers to drugs administered by health care providers at the site of care, whereas drug *order* refers to prescriptions for drugs used at home.

[d]Task 1 was applied to the full data sets, including 7366 and 11,386 patients in Data sets A and B, respectively.

[e]Task 3 was applied for patients with the first advanced HNC diagnosis on or after January 1, 2015, including 4808 and 4003 patients in Data sets A and B, respectively.

[f]N/A: not applicable.

[g]Tasks 6 and 9 were not conducted because they apply to an oral target SACT.

[h]Information about the number of refills, days supply, or alternative data elements was not available in Data set B.

[i]The total number of drug order records in Data set A (3241) and Data set B (8380) was used as the denominator in task 8.

We determined that 4808 (65.27%) of the 7366 patients in Data set A and 4003 (35.16%) of the 11,386 patients in Data set B had a first advanced HNC diagnosis on or after January 1, 2015, the timeline we applied for the study index date as it covered the key diagnostic and therapeutic timeline of the target SACT (first approved in 2016). A total of 1200 (24.96%) of the 4808 patients meeting this timeline in Data set A and 237 (5.92%) of the 4003 patients meeting this timeline in Data set B had a record of receiving the target SACT (Table 3, task 3).

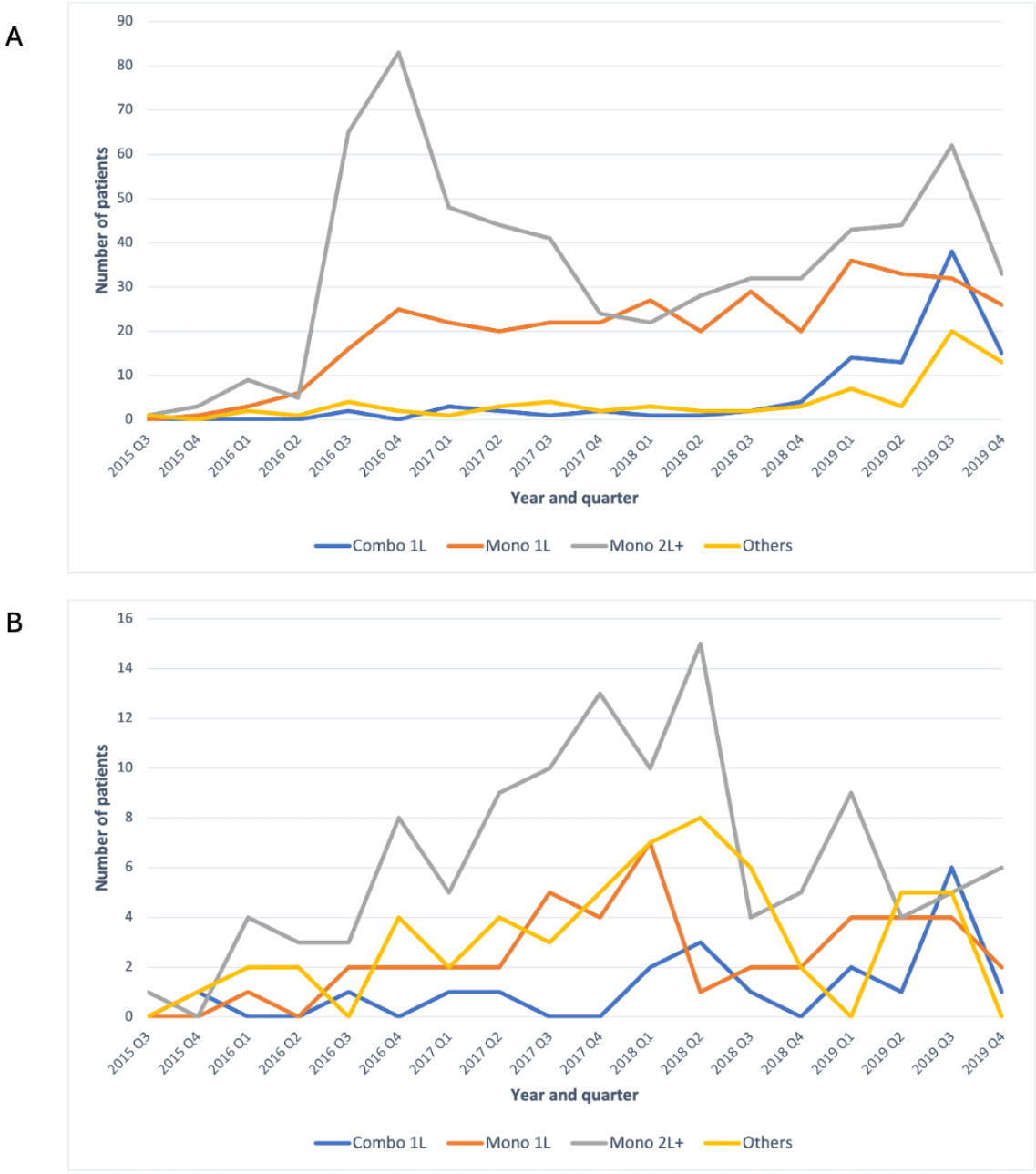
The median length of the gap between target SACT administrations was 21 days in both the data sets, which aligned with the expected dose schedule for the target SACT (Table 3, task 5). However, the range of the gap was considerably shorter in Data set A (1-113 d) than in Data set B (1-824 d), suggesting incomplete target SACT administration records in Data set B.

For the oral SACT records, Data set A included the number of refills and a flag for canceled medication orders, whereas Data set B did not provide refill information (Table 3, task 7). This could impact the accuracy of calculating rwTTD for an orally dispensed SACT because the drug orders for patients remaining on treatment through refills would not be recorded in the database.

LOT Data Checks

The 2 data sets differed in terms of the target SACT LOT distribution over time. The cumulative frequency of target SACT initiation, including as monotherapy or combination therapy and in any LOT, tended to be greater in later years in Data set A, peaking in the third quarter (Q3) of 2019, than in Data set B, peaking in the first and second quarters of 2018 (Figure 1, task 10). In Data set A, a greater frequency of target SACT initiation as second-line or later monotherapy was consistent with approval timing in this setting (2016), which preceded the first-line approvals (2019). The later time points for second-line or later monotherapy initiation in Data set B suggest the possibility of longer data lags than for Data set A.

Figure 1. Task 10: number of patients initiating the target systemic anticancer therapy (SACT) by year and quarter in (A) Data set A and (B) Data set B. Note: Y-axis heights in panels A and B differ but were selected to best depict the patient numbers in Data sets A and B. 1L: first-line therapy; 2L+: second-line or later therapy; combo: target SACT in any combination therapy (approved or not approved); mono: target SACT monotherapy; Q1: first quarter; Q2: second quarter; Q3: third quarter; Q4: fourth quarter.



In both data sets, we observed the inclusion of patients who initiated the target SACT therapy before the applicable first-line or second-line or later US Food and Drug Administration approval dates. We believe that these are true real-world findings, which do not always correspond to recommended or approved indications, rather than data quality issues.

In Data set B, only 40.3% (4589/11386) of patients had SACT LOT records (Table 4, task 11), which coincides with the finding of lower-than-expected SACT drug administration and order rates (Table 3, task 1).

Table 4. LOT[a] rules for SACT[b] and mortality information.

Task	Data set A	Data set B
Task 10: number of patients initiating the target SACT by year and quarter	Figure 1A	Figure 1B
Task 11: completeness of LOT information, n (%)[c]		
Patients with complete line number	5594 (75.94)	4589 (40.3)
Patients with complete line name	5594 (75.94)	N/A[d,e]
Patients with complete line start date	5594 (75.94)	4589 (40.3)
Patients with complete line end date	5594 (75.94)	N/A[e]
Task 12: patients for whom the first LOT number after the advanced HNC[f] diagnosis date was not 1, n (%)[c]	0 (0)	434 (3.81)
Task 13: distribution of LOT number at target SACT initiation		
Patients who received the target SACT, n	1200	237
Line 1, n (%)	481 (40.08)	65 (27.43)
Line 2, n (%)	486 (40.5)	92 (38.82)
Line 3, n (%)	161 (13.42)	54 (22.78)
Line 4, n (%)	46 (3.83)	16 (6.75)
Line 5, n (%)	13 (1.83)	10 (4.22)
Lines 6-10, n (%)	13 (1.83)	0 (0)
Task 14: use of target SACT in 1L[g] before approval date		
1L monotherapy		
Patients who received target SACT, % (n/N)	78.37 (377/481)	67.69 (44/65)
First administration date[h] in database	July 15, 2015	November 10, 2014
Cutoff date for the earliest 5% receipt	August 29, 2016	February 4, 2016
Cutoff date for the earliest 10% receipt	November 3, 2016	September 23, 2016
Cutoff date for the earliest 25% receipt	June 28, 2017	April 27, 2017
Approved 1L combination		
Patients who received the target SACT in approved 1L combination, % (n/N)	7.69 (37/481)	6.15 (4/65)
First administration date in database	December 18, 2018	July 1, 2019
Cutoff date for the earliest 5% receipt	February 18, 2019	N/A[i]
Cutoff date for the earliest 10% receipt	April 2, 2019	N/A[i]
Cutoff date for the earliest 25% receipt	July 9, 2019	N/A[i]
Task 15: patients with death record, n (%)[c]	4695 (63.74)	3531 (31)
Task 16: patients with multiple death records on different dates, n (%)	0 (0)	N/A[j]
Task 17: patients with clinical records showing health care activity after death date (n=4695), n (%)	1497 (31.88)	N/A[j]
≥1 d after date of death	1436 (30.59)	N/A[j]
≥3 d after date of death	1290 (27.48)	N/A[j]
≥7 d after date of death	1002 (21.34)	N/A[j]
≥30 d after date of death	79 (1.68)	N/A[j]

[a] LOT: line of therapy.

[b]SACT: systemic anticancer therapy.

[c]Tasks 11, 12, and 15 were applied to the full data sets, including 7366 and 11,386 patients in Data sets A and B, respectively.

[d]N/A: not applicable.

[e]Line name and line end date were not available in Data set B.

[f]HNC: head and neck cancer.

[g]1L: first line of therapy after the advanced HNC diagnosis date.

[h]Dates are written as month/day/year.

[i]Not calculated as only 4 patients received the target SACT in a 1L combination LOT.

[j]Only the year of death was available in Data set B.

The LOT start date in both data sets included year, month, and day, and the minimum LOT number started from 1 (first line) after the earliest advanced HNC diagnosis date for all but 3.81% (434/11386) of the patients in Data set B (Table 4, task 12). A line number other than 1 after the advanced HNC diagnosis date suggests that either a definition different from the commonly used definition [29,30] was used or that there was an earlier advanced HNC diagnosis date that was not documented.

In Data set A, 40.08% (481/1200) of patients received the target SACT in first-line therapy and 59.91% (719/1200) in second-line or later therapy, including 13.42% (161/1200) in third-line therapy (Table 4, task 13). In Data set B, 27.4% (65/237) of patients received the target SACT in first-line therapy, and 72.6% (172/237) received it in the second-line or later therapy, with frequent third-line receipt (54/237, 22.8%). Therefore, LOT rules may have been applied differently in Data set A and Data set B.

Complete information about the start date of first-line target SACT drug administration (as both monotherapy and combination therapy) was available for 377+37=414 (86.1%) of 481 patients in Data set A and 44+4=48 (74%) of 65 patients in Data set B (Table 4, task 14). In Data set A, first-line target SACT monotherapy was initiated for the first time in 2015, and approximately 5% of first-line monotherapy initiation dates fell on or before 2016, when the target SACT was approved for second-line or later therapy. Target SACT in combination therapy was first initiated in late 2018, with approximately 25% of the initiation dates falling before the start of Q3 in 2019, shortly after the approval of first-line combination therapy. In Data set B, first-line target SACT monotherapy initiation was first recorded in the fourth quarter in 2014, earlier than in Data set A, and close to 10% of initiation dates occurred before the end of Q3 in 2016. Instead, the approved target SACT combination therapy was first initiated at the start of Q3 in 2019, in line with the approval date for this indication.

Mortality Data

Among 7366 and 11,386 patients in Data sets A and B, 4695 (63.74%) and 3531 (31%), respectively, had a recorded date of death (Table 4, task 15); and 4427 (60%) and 3093 (27%) patients, respectively, had death records within 3 years after the date of advanced HNC diagnosis. These percentage differences indicate that Data set B may have incomplete mortality records (or a high loss to follow-up).

In Data set A, one-third of patients (1497/4695, 31.88%) with a recorded date of death had clinical records recorded after the death date (Table 4, task 17), with a median of 11 days from the death date to the last activity date. Thus, clinical records could be entered into the health information system after the reported death date, but extreme values (eg, >30 d after the death date) might indicate integrity issues in collecting mortality data. This information was not available for Data set B, in which the dates of death were recorded only by year.

Follow-Up Data

In Data set A, most patients (5840/7366, 79.28% to 7269/7366, 99.86%) had recorded data for diagnosis, drug records, laboratory results, facility visits, and vital sign measurements (Table 5, task 18). Similarly, in the subset of 7754 patients in Data set B whose advanced HNC diagnosis date was on or after January 1, 2011, the earliest date in Data set A, these data categories were also recorded for most patients (6123/7754, 78.97% to 6893/7754, 88.7%). Records of medical procedures not related to drug administration and genomic testing were not available in Data set A, which could result in inaccurate estimates of follow-up times.

Table 5. Unique number of patients and patient-date pairs after the advanced HNC[a] diagnosis date (task 18): follow-up data for patients with advanced HNC diagnosis on or after January 1, 2011.

Variable	Data set A (n=7366)			Data set B (n=7754)		
	Value, n (%)	Unique patient-date pairs, n	Pairs per patient, n	Value, n (%)	Unique patient-date pairs, n	Pairs per patient, n
Diagnosis	6567 (89.15)	60,178	9.2	6893 (88.9)	370,671	53.8
Drug records[b]	5840 (79.28)	113,948	19.5	6802 (87.72)	269,225	39.6
Laboratory records	6860 (93.13)	179,177	26.1	6403 (82.58)	147,314	23
Facility visit	7269 (98.68)	274,714	37.8	6838 (88.19)	392,175	57.4
Vital sign measurements	7254 (98.48)	233,623	32.2	6123 (78.97)	217,797	35.6
Nondrug medical procedure	N/A[c]	N/A	N/A	6740 (86.92)	390,556	57.9
Genomic test	N/A	N/A	N/A	118 (1.52)	208	1
Biomarker test	440 (5.97)	469	1.1	N/A	N/A	N/A
ECOG PS[d]	5416 (73.53)	100,607	17.7	N/A	N/A	N/A

[a]HNC: head and neck cancer.

[b]Any drug, not just systemic anticancer therapies.

[c]N/A: not applicable.

[d]ECOG PS: Eastern Cooperative Oncology group performance status.

The median frequency of visits (normalized by length between first and last target SACT administration) for patients who received the target SACT was somewhat less in Data set A, varying from 0.05 to 0.12, depending on treatment line, than in Data set B, in which it varied from 0.14 to 0.18 (Table 6, task 19). This might indicate that more clinical activities were recorded in Data set B during treatment.

Table 6. Follow-up data for patients with advanced HNC[a] diagnosis on or after January 1, 2011.

Task	Data set A (n=7366)			Data set B (n=7754)		
	Value, n	Value, median (IQR; range)	Value, mean (SD)	Value, n	Value, median (IQR; range)	Value, mean (SD)
Task 19: frequency of visits during target SACT[b,c]						
1L[d] combination therapy	101	0.11 (0.07-0.16; 0.02-0.33)	0.13 (0.07)	19	0.17 (0.11-0.24; 0-0.36)	0.17 (0.09)
1L monotherapy	358	0.05 (0.05-0.08; 0.01-0.50)	0.07 (0.05)	44	0.18 (0.09-0.22; 0-0.48)	0.16 (0.11)
2L+[e] monotherapy	634	0.06 (0.05-0.10; 0.01-0.95)	0.08 (0.06)	106	0.13 (0.07-0.25; 0-0.75)	0.17 (0.14)
All other	104	0.12 (0.09-0.17; 0.02-0.48)	0.14 (0.08)	76	0.14 (0.09-0.21; 0-1.3)	0.17 (0.17)
Task 20: for patients still alive, gap (in d) from the last target SACT administration and last visit[f]	708	28 (6-187; 0-1118)	128 (199)	167	70 (29-223; 0-1755)	159 (215)

[a]HNC: head and neck cancer.

[b]SACT: systemic anticancer therapy.

[c]Frequency defined as number of visits between the first and last target SACT administration dates within the same LOT number and name, divided by number of days between the last and first target SACT administration.

[d]1L: first line of therapy after the advanced HNC diagnosis date.

[e]2L+: second-line or later therapy.

[f]Limited to patients who (1) were still alive ≥180 days after last receipt of target SACT and (2) received last dose of target SACT ≥180 days before data cutoff on November 25, 2019 (thus on or before May 29, 2019).

Discussion

Principal Findings

This study identified 20 data quality assessment tasks for the use case of estimating the rwTTD of an SACT. By executing the 18 tasks pertinent to the intravenously administered target SACT, we demonstrated that the UReQA framework for the rwTTD use case can be implemented to generate descriptive summary statistics and charts. These visualizations provide additional insights into the relevance and quality of 2 US EHR-based oncology RWD. The approach is generalizable to implement for other SACT and databases.

Both data sets in the evaluation provided all the required data elements; however, verification checks revealed that Data set B might not be suitable for analyzing rwTTD for the target SACT because (1) the large decrease in patient receiving the target SACT in recent years suggests longer lags in incorporating the most recent data and (2) the completeness and plausibility issues in the SACT, LOT, and mortality data could cause faulty determination of treatment discontinuation date and status of censoring.

The fact that Data set B included a lower percentage of patients receiving the target SACT (237/4003, 5.9% vs 1200/4808, 24.96% in Data set A) limited the utility of the data for determining the rwTTD. This finding highlights the need and importance of conducting a rigorous and use case–specific data quality assessment in the planning stage of RWD studies. In addition, for Data set B, findings of extremely low and high gaps between target SACT administration dates would warrant further investigation of each patient's trajectory to verify the specific data quality issue before taking proper data quality improvement actions such as removing the patient or the SACT record as outliers.

Limitations

This study has several limitations that require further discussion. First, adequately assessing the reasons for missingness across different RWD sources is challenging. In particular, the data feeds and capture of elements across different data sources are variable. A lack of transparency and consistency means that different RWD sources are often not fully interoperable [38]. In this study, we applied cohort attrition steps to align populations represented in the 2 data sets and imputed the LOT end date and LOT name that were missing in Data set B. However, a major remaining roadblock was the vendor's privacy-preserving aggregation, which does not allow data sources to be adequately reviewed on more granular level to understand the reason behind missing data, data quality issues, or data discrepancies.

Second, the implementation of data quality checks for new RWD sources, especially for those with data table structures that differ from those of prior data sets, requires customization and reconfiguration that are often time consuming. We are developing a data dashboard tool that can accelerate this process for both raw data and a common data model such as that of the Observational Health Data Sciences and Informatics [17,18].

Third, use case–specific data quality assessment checks often provide only a limited view of the comparative validity of the RWD under consideration, particularly when a well-recognized gold standard is absent. The paucity of data often limits an effective comparison with the distribution of key data elements in the general population (external validity). In this study, we set a priori metrics for these checks by using domain knowledge such as HNC prevalence [33] and regulatory approval timelines. It would be interesting for future studies to validate and update these metrics.

Comparison With Prior Work

Prior studies have evaluated rwTTD, also known as the duration of therapy and real-world time on treatment, for immuno-oncology agents used in treating recurrent or metastatic HNC [39], advanced non-small cell lung cancer [28,40-42], and other solid cancers [42]. In contrast to this study, these studies drew on research-ready databases (as would be identified in the preassessment step of UReQA [3]), and the actions taken to ensure RWD fitness and quality were limited to aligning patient eligibility criteria (the cohort definition step of UReQA [3]).

New use cases can be created for other medication-related outcomes or therapeutic areas by following the first 3 steps of implementing the rwTTD use case in this study. In addition, the data quality checks that we identified and created for the rwTTD use case can be used for other types of use cases. For example, checks on medication identification and dates can also be used to evaluate the fitness of RWD sources for studying medication adherence. The checks on mortality and follow-up visits could validate the applicability of an RWD source for survival analyses.

Future Work

We selected 2 US EHR-based oncology databases to implement the UReQA use case of rwTTD. These were the only 2 databases the research team had access to that provided both oncology treatment and LOT information during the time of study execution. Each database may have its own bias in representing the overall advanced HNC population in the United States. Future work could implement (1) evaluation of more US EHR-based oncology databases to bring more impactful findings and (2) investigating the associations between rwTTD calculation and quantitative data quality assessment for various medications of interest and cancer types.

Conclusions

The fit-for-purpose quality assessment demonstrated the high level of variability in quality of the 2 real-world data sets for estimating the rwTTD of an SACT for advanced HNC. This study illustrates the application and value of use case–specific data assessment tasks in identifying high-quality RWD for research studies. The data quality specifications supporting this comprehensive use case can be expanded to other use cases in oncology outcomes research. Incorporating such comprehensive data quality assessment could help the study team select the most suitable database in the planning stage of a real-world evidence study. In addition, understanding data quality concerns particularly relevant to research questions can provide additional insights for properly preparing data in full study execution.

Acknowledgments

This work was supported by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States. Medical writing and editorial assistance were provided by Elizabeth V Hillyer, DVM (freelance). This assistance was funded by Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, New Jersey, United States.

Data Availability

The data sets generated during and/or analyzed during this study are not publicly available as they were data vendors' proprietary assets provided to the study team under commercial licenses but are available from the corresponding author on reasonable request and permission from the data vendor.

In addition, we cannot disclose the identities of the data vendors, as doing so would inevitably promote the business of 1 data vendor and may violate data use agreements.

Authors' Contributions

BR, AS, KD, and SC conceptualized and designed the study. BR and AS contributed to data acquisition and data analysis. BR, AS, KD, SC, LY, and SK contributed to the interpretation of results. BR and AS drafted the manuscript. BR, AS, KD, SC, LY, and SK contributed to manuscript revision. All authors approved the publication of the manuscript.

Conflicts of Interest

BR, KD, and SK report employment with Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, NJ, United States, and stock ownership of Merck & Co, Inc, Rahway, NJ, United States. AS reports employment with Real World Evidence, Epidemiology, Medical Affairs and Value Statistics (REM) Data Science department, Jazz Pharmaceutical. SC reports employment with ConcertAI. LY reports employment with and ownership of Polygon Health Analytics LLC. AS, SC, and LY were employees of Merck Sharp & Dohme LLC, a subsidiary of Merck & Co, Inc, Rahway, NJ, United States, when they worked on this study.

Multimedia Appendix 1
Data checks comprising 20 tasks assessing conformance, completeness, or plausibility.
[DOCX File , 28 KB - medinform_v12i1e47744_app1.docx]

References

1. Khozin S, Blumenthal GM, Pazdur R. Real-world data for clinical evidence generation in oncology. J Natl Cancer Inst 2017 Nov 01;109(11):1-5. [doi: 10.1093/jnci/djx187] [Medline: 29059439]
2. Miksad RA, Samant MK, Sarkar S, Abernethy AP. Small but mighty: the use of real-world evidence to inform precision medicine. Clin Pharmacol Ther 2019 Jul;106(1):87-90 [FREE Full text] [doi: 10.1002/cpt.1466] [Medline: 31112289]
3. Desai KD, Chandwani S, Ru B, Reynolds MW, Christian JB, Estiri H. Fit-for-purpose real-world data assessments in oncology: a call for cross-stakeholder collaboration. Value Outcomes Spotlight 2021 Jun;24:S25 [FREE Full text] [doi: 10.1016/j.jval.2021.04.129]
4. Dagenais S, Russo L, Madsen A, Webster J, Becnel L. Use of real-world evidence to drive drug development strategy and inform clinical trial design. Clin Pharmacol Ther 2022 Jan;111(1):77-89 [FREE Full text] [doi: 10.1002/cpt.2480] [Medline: 34839524]
5. Lakdawalla DN, Shafrin J, Hou N, Peneva D, Vine S, Park J, et al. Predicting real-world effectiveness of cancer therapies using overall survival and progression-free survival from clinical trials: empirical evidence for the ASCO value framework. Value Health 2017;20(7):866-875 [FREE Full text] [doi: 10.1016/j.jval.2017.04.003] [Medline: 28712615]
6. Framework for FDA's real-world evidence program. U.S. Food & Drug Administration (FDA). 2018. URL: https://www.fda.gov/media/120060/download [accessed 2023-03-10]
7. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. U.S. Department of Health and Human Services. 2021. URL: https://www.fda.gov/media/152503/download [accessed 2023-03-10]
8. Submitting documents using real-world data and real-world evidence to FDA for drug and biological products: guidance for industry. U.S. Food & Drug Administration (FDA). 2022 Sep. URL: https://www.regulations.gov/document/FDA-2019-D-1263-0014 [accessed 2023-03-10]
9. Snapshot: healthcare data ecosystem (2023). DATAVANT. URL: https://datavant.com/health-data-ecosystem/ [accessed 2023-10-13]

10. 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]
11. Assessing data quality for healthcare systems data used in clinical research. NIH Pragmatic Trials Collaboratory. 2014. URL: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf [accessed 2023-03-10]
12. Franklin JM, Liaw KL, Iyasu S, Critchlow CW, Dreyer NA. Real-world evidence to support regulatory decision making: new or expanded medical product indications. Pharmacoepidemiol Drug Saf 2021 Jun;30(6):685-693. [doi: 10.1002/pds.5222] [Medline: 33675248]
13. 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]
14. Mahendraratnam N, Silcox C, Mercon K, Kroetsch A, Romine M, Harrison N, et al. Determining real-world data?s fitness for use and the role of reliability: Duke-Margolis Center for Health Policy. Duke Margolis Center for Health Policy. 2019. URL: https://healthpolicy.duke.edu/publications/determining-real-world-datas-fitness-use-and-role-reliability [accessed 2023-03-10]
15. Callahan TJ, Bauck AE, Bertoch D, Brown J, Khare R, Ryan PB, et al. A comparison of data quality assessment checks in six data sharing networks. EGEMS (Wash DC) 2017 Jun 12;5(1):8 [FREE Full text] [doi: 10.5334/egems.223] [Medline: 29881733]
16. Reynolds MW, Bourke A, Dreyer NA. Considerations when evaluating real-world data quality in the context of fitness for purpose. Pharmacoepidemiol Drug Saf 2020 Oct;29(10):1316-1318 [FREE Full text] [doi: 10.1002/pds.5010] [Medline: 32374042]
17. OHDSI: data quality dashboard. GitHub. URL: https://github.com/OHDSI/DataQualityDashboard [accessed 2023-03-10]
18. Home page. Observational Health Data Sciences and Informatics (OHDSI). URL: https://www.ohdsi.org/ [accessed 2023-03-10]
19. PEDSnet/data quality analysis. GitHub. URL: https://github.com/PEDSnet/Data-Quality-Analysis/blob/master/Data/DQACatalog/DQA_Check_Type_Inventory.csv [accessed 2023-03-10]
20. Home page. The National Patient-Centered Clinical Research Network (PCORnet). URL: https://pcornet.org/ [accessed 2023-03-10]
21. DQUEEN v 0.5 (data QUality assEssmENt and managing tool). GitHub. URL: https://github.com/ABMI/DQUEEN_OMOP_CDM_Version [accessed 2023-03-10]
22. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. BMC Med Inform Decis Mak 2019 Jul 23;19(1):142 [FREE Full text] [doi: 10.1186/s12911-019-0852-6] [Medline: 31337390]
23. 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]
24. Estiri H, Klann JG, Weiler SR, Alema-Mensah E, Joseph Applegate R, Lozinski G, et al. A federated EHR network data completeness tracking system. J Am Med Inform Assoc 2019 Jul 01;26(7):637-645 [FREE Full text] [doi: 10.1093/jamia/ocz014] [Medline: 30925587]
25. Huser V. Facilitating analysis of measurements data though stricter model conventions: exploring units variability across sites. Observational Health Data Sciences and Informatics. 2017. URL: https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=resources:huser-2017-ohdsi-symp-units.pdf [accessed 2023-03-10]
26. Blumenthal GM, Gong Y, Kehl K, Mishra-Kalyani P, Goldberg KB, Khozin S, et al. Analysis of time-to-treatment discontinuation of targeted therapy, immunotherapy, and chemotherapy in clinical trials of patients with non-small-cell lung cancer. Ann Oncol 2019 May 01;30(5):830-838 [FREE Full text] [doi: 10.1093/annonc/mdz060] [Medline: 30796424]
27. Establishing a framework to evaluate real-world endpoints. Friends of Cancer Research. 2018. URL: https://friendsofcancerresearch.org/wp-content/uploads/RWE_FINAL-7.6.18_1.pdf [accessed 2023-03-10]
28. Stewart M, Norden AD, Dreyer N, Henk HJ, Abernethy AP, Chrischilles E, et al. An exploratory analysis of real-world end points for assessing outcomes among immunotherapy-treated patients with advanced non-small-cell lung cancer. JCO Clin Cancer Inform 2019 Jul;3:1-15 [FREE Full text] [doi: 10.1200/CCI.18.00155] [Medline: 31335166]
29. Meng W, Ou W, Chandwani S, Chen X, Black W, Cai Z. Temporal phenotyping by mining healthcare data to derive lines of therapy for cancer. J Biomed Inform 2019 Dec;100:103335 [FREE Full text] [doi: 10.1016/j.jbi.2019.103335] [Medline: 31689549]
30. Meng W, Mosesso KM, Lane KA, Roberts AR, Griffith A, Ou W, et al. An automated line-of-therapy algorithm for adults with metastatic non-small cell lung cancer: validation study using blinded manual chart review. JMIR Med Inform 2021 Oct 12;9(10):e29017 [FREE Full text] [doi: 10.2196/29017] [Medline: 34636730]
31. RxNorm. NIH National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/rxnorm/index.html [accessed 2023-03-10]
32. RxNav-in-a-box. NIH National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/RxNav/applications/RxNav-in-a-Box.html [accessed 2023-03-10]

33. Cancer stat facts: oral cavity and pharynx cancer. National Cancer Institute: Surveillance, Epidemiology, and End Results Program (SEER). 2022. URL: https://seer.cancer.gov/statfacts/html/oralcav.html [accessed 2023-03-10]
34. Nadler E, Joo S, Boyd M, Black-Shinn J, Chirovsky D. Treatment patterns and outcomes among patients with recurrent/metastatic squamous cell carcinoma of the head and neck. Future Oncol 2019 Mar;15(7):739-751. [doi: 10.2217/fon-2018-0572] [Medline: 30511880]
35. Grünwald V, Chirovsky D, Cheung WY, Bertolini F, Ahn MJ, Yang MH, et al. Global treatment patterns and outcomes among patients with recurrent and/or metastatic head and neck squamous cell carcinoma: results of the GLANCE H and N study. Oral Oncol 2020 Mar;102:104526 [FREE Full text] [doi: 10.1016/j.oraloncology.2019.104526] [Medline: 31978755]
36. Mody MD, Rocco JW, Yom SS, Haddad RI, Saba NF. Head and neck cancer. Lancet 2021 Dec 18;398(10318):2289-2299. [doi: 10.1016/S0140-6736(21)01550-6] [Medline: 34562395]
37. Mourad M, Jetmore T, Jategaonkar AA, Moubayed S, Moshier E, Urken ML. Epidemiological trends of head and neck cancer in the united states: a SEER population study. J Oral Maxillofac Surg 2017 Dec;75(12):2562-2572 [FREE Full text] [doi: 10.1016/j.joms.2017.05.008] [Medline: 28618252]
38. Penberthy LT, Rivera DR, Lund JL, Bruno MA, Meyer AM. An overview of real-world data sources for oncology and considerations for research. CA Cancer J Clin 2022 May;72(3):287-300 [FREE Full text] [doi: 10.3322/caac.21714] [Medline: 34964981]
39. Ramakrishnan K, Liu Z, Baxi S, Chandwani S, Joo S, Chirovsky D. Real-world time on treatment with immuno-oncology therapy in recurrent/metastatic head and neck squamous cell carcinoma. Future Oncol 2021 Aug;17(23):3037-3050 [FREE Full text] [doi: 10.2217/fon-2021-0360] [Medline: 34044594]
40. Waterhouse D, Lam J, Betts KA, Yin L, Gao S, Yuan Y, et al. Real-world outcomes of immunotherapy-based regimens in first-line advanced non-small cell lung cancer. Lung Cancer 2021 Jun;156:41-49 [FREE Full text] [doi: 10.1016/j.lungcan.2021.04.007] [Medline: 33894493]
41. Horvat P, Gray CM, Lambova A, Christian JB, Lasiter L, Stewart M, et al. Comparing findings from a friends of cancer research exploratory analysis of real-world end points with the cancer analysis system in England. JCO Clin Cancer Inform 2021 Dec;5:1155-1168 [FREE Full text] [doi: 10.1200/CCI.21.00013] [Medline: 34860576]
42. Torres AZ, Nussbaum NC, Parrinello CM, Bourla AB, Bowser BE, Wagner S, et al. Analysis of a real-world progression variable and related endpoints for patients with five different cancer types. Adv Ther 2022 Jun;39(6):2831-2849 [FREE Full text] [doi: 10.1007/s12325-022-02091-8] [Medline: 35430670]

Abbreviations

EHR: electronic health record
HNC: head and neck cancer
ICD: International Classification of Diseases
LOT: line of therapy
Q3: third quarter
RWD: real-world data
rwTTD: real-world time to treatment discontinuation
SACT: systemic anticancer therapy
UReQA: Use Case Specific Relevance and Quality Assessment

Edited by Q Chen; submitted 30.03.23; peer-reviewed by HJ Kim, S Setia, T Royce; comments to author 30.06.23; revised version received 30.11.23; accepted 14.01.24; published 06.03.24.

Please cite as:
Ru B, Sillah A, Desai K, Chandwani S, Yao L, Kothari S
Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation Use Case: Implementation and Evaluation Study
JMIR Med Inform 2024;12:e47744
URL: https://medinform.jmir.org/2024/1/e47744
doi:10.2196/47744
PMID:38446504

Knowledge Graph for Breast Cancer Prevention and Treatment: Literature-Based Data Analysis Study

Shuyan Jin[1], MPH; Haobin Liang[2], MSc; Wenxia Zhang[1], PhD; Huan Li[1], MM

1

2

Corresponding Author:
Shuyan Jin, MPH

Abstract

Background: The incidence of breast cancer has remained high and continues to rise since the 21st century. Consequently, there has been a significant increase in research efforts focused on breast cancer prevention and treatment. Despite the extensive body of literature available on this subject, systematic integration is lacking. To address this issue, knowledge graphs have emerged as a valuable tool. By harnessing their powerful knowledge integration capabilities, knowledge graphs offer a comprehensive and structured approach to understanding breast cancer prevention and treatment.

Objective: We aim to integrate literature data on breast cancer treatment and prevention, build a knowledge graph, and provide support for clinical decision-making.

Methods: We used Medical Subject Headings terms to search for clinical trial literature on breast cancer prevention and treatment published on PubMed between 2018 and 2022. We downloaded triplet data from the Semantic MEDLINE Database (SemMedDB) and matched them with the retrieved literature to obtain triplet data for the target articles. We visualized the triplet information using NetworkX for knowledge discovery.

Results: Within the scope of literature research in the past 5 years, malignant neoplasms appeared most frequently (587/1387, 42.3%). Pharmacotherapy (267/1387, 19.3%) was the primary treatment method, with trastuzumab (209/1805, 11.6%) being the most commonly used therapeutic drug. Through the analysis of the knowledge graph, we have discovered a complex network of relationships between treatment methods, therapeutic drugs, and preventive measures for different types of breast cancer.

Conclusions: This study constructed a knowledge graph for breast cancer prevention and treatment, which enabled the integration and knowledge discovery of relevant literature in the past 5 years. Researchers can gain insights into treatment methods, drugs, preventive knowledge regarding adverse reactions to treatment, and the associations between different knowledge domains from the graph.

(JMIR Med Inform 2024;12:e52210) doi:10.2196/52210

KEYWORDS

knowledge graph; breast cancer; treatment; prevention; adverse reaction

Introduction

Breast cancer is the most common malignant tumor in women worldwide, with a reported death toll exceeding 600,000 in 2018 alone [1]. Breast cancer has emerged as the most prevalent cancer and a primary cause of mortality among women. The global incidence of new cases of female breast cancer witnessed a sharp increase from 1.05 million in 2000 to 2.09 million in 2018 [2]. In 2020, global cancer burden data revealed that new breast cancer cases reached 2.26 million, constituting 11.7% of all newly diagnosed cancer cases worldwide. The newly reported mortality cases numbered 0.68 million, representing 6.9% of global newly reported deaths [3]. Factors such as old age, young age at menarche, family history of breast cancer, smoking, and drinking alcohol increase the risk of breast cancer [4-6]. On the contrary, regular physical exercise; breastfeeding; regular work and rest; and intake of fruits, vegetables, whole grains, and dietary fiber can appropriately reduce the risk of breast cancer [7]. Various treatment methods are used for patients with breast cancer, including surgery, radiation therapy, endocrine therapy, chemotherapy, and targeted therapy. So far, most countries have primarily focused on population education for breast cancer prevention, including encouraging increased physical activity, controlling BMI, and limiting alcohol intake [8]. Despite the increasing number of research literature, a large amount of literature on breast cancer prevention and treatment has not been systematically integrated. Knowledge graph technology allows for the independent connection and integration of disparate literature, resulting in a more comprehensive and cohesive knowledge framework.

Knowledge Graph is a knowledge repository proposed by Google in 2012 to enhance the functionality of search engines.

It describes concepts and their relationships in the real world using triplets in the form of entity-relation-entity [9]. Knowledge graphs can integrate information from diverse sources and domains, including text, databases, and web pages, and intricately interlink them. These integrations serve to mitigate information silos, fostering the establishment of a more comprehensive knowledge framework. Knowledge graphs have been widely used in various fields, such as medicine, network security, journalism, finance, and education [10]. Knowledge graphs in the biomedical domain have applications in studies related to disease associations [11], genomics [12], drug interactions [13], and support for physicians in formulating individualized treatment regimens [14]. At present, there are well-established knowledge graphs, including DisGeNET [15], which integrate information on the associations between genes and diseases; DrugBank [16], a comprehensive bioinformatics and cheminformatics knowledge base; and ClinVar [17], a compilation of genetic variation information from diverse laboratories worldwide. One study extracted breast cancer–related features from Chinese breast cancer mammography reports and built a knowledge graph for diagnosing breast cancer by combining diagnosis and treatment guidelines and insights from clinical experts [18]. Another study integrated triples from clinical guidelines, medical encyclopedias, and electronic medical records to build a breast cancer knowledge graph [19]. Despite a small number of scholars having constructed knowledge graphs for breast cancer, the varied emphases and diverse data sources employed render their applicability limited. A knowledge graph specifically focused on the prevention and treatment of breast cancer has not been constructed at present. Therefore, this study primarily collects information related to the prevention and treatment of breast cancer to construct a knowledge graph.

In the biomedical field, there are already mature tools (eg, SemRep) for extracting knowledge from medical texts. SemRep is a natural language processing program based on the Unified Medical Language System (UMLS), which performs operations such as text tokenization, syntactic analysis, part-of-speech disambiguation, phrase mapping, semantic predicate normalization, and syntactic constraints [20]. It extracts entities and relationships from biomedical texts and outputs triplets stored in the Semantic MEDLINE Database (SemMedDB) [21]. SemMedDB currently encompasses details on approximately 96.3 million predications derived from all PubMed citations (around 29.1 million citations) and serves as the foundation for the Semantic MEDLINE application [22]. We downloaded the entity and relationship data provided by SemMedDB. NetworkX is an open-source library for Python, primarily designed for creating, analyzing, and visualizing complex network structures. NetworkX plays a significant role in knowledge visualization, facilitating users in intuitively presenting and comprehending intricate knowledge graphs or network data.

Methods

Ethics Approval

This study was approved by the Board of Medical Ethics Committee of Shenzhen Maternal and Child Health Hospital (SFYLS[2022]003).

Data Source

We conducted a search on PubMed using Medical Subject Headings terms “breast cancer,” “prevention,” and “treatment,” covering the period from January 1, 2018, to December 31, 2022, and the study type was clinical trials. A total of 3589 articles were retrieved. We obtained the entity and relationship data from SemMedDB.

Data Processing and Construction of Knowledge Graph

We matched the PMIDs of the retrieved articles with the database and extracted the corresponding triplet information. We initially obtained 33,060 Subject-Predicate-Object (SPO) triplets of data.

Next, we made improvements according to the SPO cleaning principles proposed by Fiszman et al [9] (ie, relevance, connectivity, novelty, and significance). We combined them with expert manual screening to ensure that the selected SPO triplets have a higher relevance. In the improved process, we did not predefine semantic patterns. Instead, we used a series of cleaning operations to select core SPO triplets and connected SPO triplets, eliminating SPO triplets lacking specific information and those that appeared only once in the frequency. The specific process is as follows:

1. In the same article, there may be repeated occurrences of identical SPO triplets. To maintain equal contribution from each article, we counted the repeated SPO triplets once within the same article.
2. To ensure statistical reliability, we calculated the occurrence frequency of each SPO triplet across different articles. SPO triplets with low occurrence frequencies may lack statistical significance. Therefore, we filtered SPO triplets with frequencies greater than or equal to 2.
3. Based on expert domain knowledge, we manually screened the selected SPO triplets with frequencies greater than or equal to 2 to identify those of research value.

Finally, we obtained 25,449 SPO triplets data. We imported the filtered SPO triplets information into the NetworkX for visual analysis to explore knowledge and information related to breast cancer prevention and treatment.

All analyses were conducted in a Python program (version 3.11.3; Python Software Foundation), primarily using Pandas, Matplotlib, WordCloud, and NetworkX packages [23-26].

Results

Summary of Included Literatures

A total of 3589 articles were published in 618 different journals. Among them, 191 articles were published in the same journal, while 293 journals had only 1 article published. The journals

were ranked based on the number of publications, and the top 100 journals accounted for 2631 articles, which is 73.30% of the total.

Semantic Relationships and Semantic Patterns

We mainly summarize semantic associations into 3 types: treatment and prevention, influencing or associated factors, and related diseases (Table S1 in Multimedia Appendix 1). Regarding treatment and prevention, the relationships include TREATS, ADMINISTERED_TO, USES, and PREVENTS, representing treatment drugs, surgeries, and preventive measures for breast cancer. Regarding influencing or associated factors, the relationships include ASSOCIATED_WITH, AFFECTS, and CAUSES, which represent diseases' impact and etiological factors. Regarding related diseases, the relationship COEXISTS_WITH represents the coexistence between different diseases. In the semantic patterns involving treatment (TREATS), the topp-TREATS-neop and topp-TREATS-podg have appeared over 1000 times.

Summary of SPO Triples

In terms of breast tumors, malignant neoplasms had the highest frequency, accounting for 42.3% (587/1387) of the total, followed by triple-negative breast neoplasms (56/1387, 4%) and human epidermal growth factor receptor 2 (*HER*2)–positive carcinoma of breast (54/1387, 4%; Table 1 and Multimedia Appendix 2).

Table . Summary of breast cancer subtypes and stages, treatment methods, and treatment drugs. The top 30 subtypes, treatment methods, and treatment drugs with higher frequencies in all data are presented for each group.

Group		Values, n (%)
Breast cancer subtypes and stages (n=1387)		
	Malignant neoplasm of breast	587 (42.3)
	Triple-negative breast neoplasms	56 (4)
	HER2[a]-positive carcinoma of breast	54 (3.9)
	Carcinoma breast stage IV	48 (3.5)
	Breast cancer metastatic	47 (3.4)
	Early-stage breast carcinoma	42 (3)
	Malignant neoplasms	31 (2.2)
	Neoplasm	30 (2.2)
	Metastatic triple-negative breast carcinoma	26 (1.9)
	High-risk cancer	24 (1.7)
	Neoplasm metastasis	21 (1.5)
	Advanced cancer	19 (1.4)
	Advanced breast carcinoma	19 (1.4)
	HER2-negative breast cancer	18 (1.3)
	Locally advanced malignant neoplasm	17 (1.2)
	Advanced malignant neoplasm	15 (1.1)
	Nonsmall cell lung carcinoma	15 (1.1)
	Noninfiltrating intraductal carcinoma	14 (1)
	Locally advanced breast cancer	13 (0.9)
	Breast cancer stage III	11 (0.8)
Treatment of breast cancer (n=1387)		
	Pharmacotherapy	267 (19.3)
	Neoadjuvant therapy	88 (6.3)
	Hormone therapy	68 (4.9)
	Chemotherapy (adjuvant)	54 (3.9)
	Therapeutic procedure	53 (3.8)
	Radiation therapy	48 (3.5)
	Treatment protocols	43 (3.1)
	Adjuvant therapy	36 (2.6)
	Breast-conserving surgery	35 (2.5)
	First-line treatment	31 (2.2)
	Single-agent therapy	27 (1.9)
	Mastectomy	27 (1.9)
	Operative surgical procedures	20 (1.4)
	Interventional procedure	16 (1.2)
	Radiotherapy (adjuvant)	14 (1)
	Excision of axillary lymph nodes group	13 (0.9)
	Combined modality therapy	12 (0.9)
	Excision	11 (0.8)
	Targeted therapy	11 (0.8)

Group		Values, n (%)
	Placebos	10 (0.7)
Drugs for breast cancer (n=1805)		
	Trastuzumab	209 (11.6)
	Capecitabine	88 (4.9)
	Paclitaxel	81 (4.5)
	Aromatase inhibitors	64 (3.5)
	Immunologic adjuvants	62 (3.4)
	Letrozole	58 (3.2)
	Bevacizumab	48 (2.7)
	Tamoxifen	40 (2.2)
	Gemcitabine	36 (2)
	Pertuzumab	36 (2)
	Fulvestrant	36 (2)
	Cyclophosphamide	32 (1.8)
	Pembrolizumab	30 (1.7)
	Docetaxel	27 (1.5)
	Taxane	27 (1.5)
	Ado-trastuzumab emtansine	22 (1.2)
	130-nm albumin-bound paclitaxel	22 (1.2)
	Carboplatin	22 (1.2)
	Eribulin	21 (1.2)
	Palbociclib	19 (1.1)
	Exemestane	19 (1.1)
	Everolimus	19 (1.1)
	Olaparib	18 (1)
	Talazoparib	17 (0.9)
	Pharmaceutical preparations	16 (0.9)
	Protein-tyrosine kinase inhibitor	15 (0.8)
	Cisplatin	14 (0.8)
	Lapatinib	14 (0.8)
	Fluorouracil	13 (0.7)
	Preservative free ingredient	13 (0.7)

[a]*HER*2: human epidermal growth factor receptor 2.

Pharmacotherapy is the most common treatment method, accounting for 19.2% (267/1387) of the overall frequency. Additionally, other high-frequency treatment modalities include neoadjuvant therapy (88/1387, 6%), hormone therapy (68/1387, 5%), adjuvant chemotherapy (54/1387, 4%), and radiation therapy (48/1387, 3%; Table 1 and Multimedia Appendix 3). In breast cancer treatment drugs, trastuzumab (209/1805, 11.6%), capecitabine (88/1805, 5%), paclitaxel (81/1805, 4%), aromatase inhibitors (64/1805, 4%), and immunologic adjuvants (62/1805, 3%) have a relatively high frequency of occurrence (Table 1 and Multimedia Appendix 4).

Breast Cancer Knowledge Graph

We visualized the SPO triples and displayed 3 subgroups: breast cancer treatment methods, therapeutic drugs, and relevant preventive measures. Figure 1 shows the relationship between different subtypes and stages of breast cancer and treatment methods. In different subtypes of breast cancer, the highest frequency is observed in malignant neoplasm of the breast, with pharmacotherapy having the highest frequency among various treatment modalities. Different subtypes simultaneously correspond to multiple treatment modalities; likewise, a single treatment modality corresponds to multiple breast cancer subtypes.

Figure 1. Relationship between different subtypes and stages of breast cancer and treatment methods. *HER2*: human epidermal growth factor receptor 2.

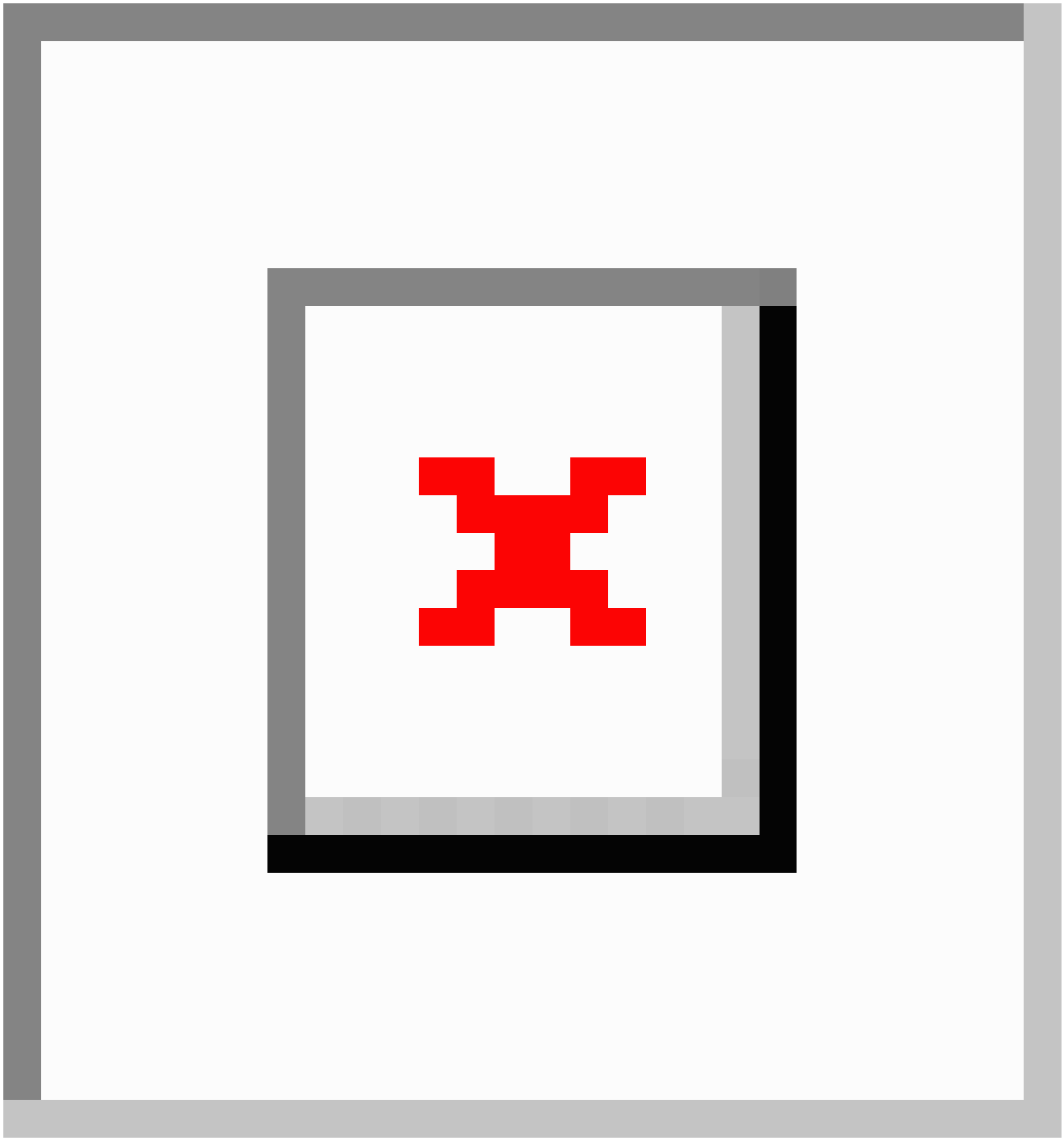


Figure 2 shows the relationship between different subtypes and stages of breast cancer and drugs. Among the therapeutic drugs for breast cancer, trastuzumab has the highest frequency and corresponds to the most types of breast cancer. Capecitabine, paclitaxel, aromatase inhibitors, and immunologic adjuvants also have relatively high frequencies. In comparison, immunologic adjuvants have the fewest connections with different types of breast cancer.

Figure 2. Relationship between different subtypes and stages of breast cancer and drugs. *HER2*: human epidermal growth factor receptor 2.

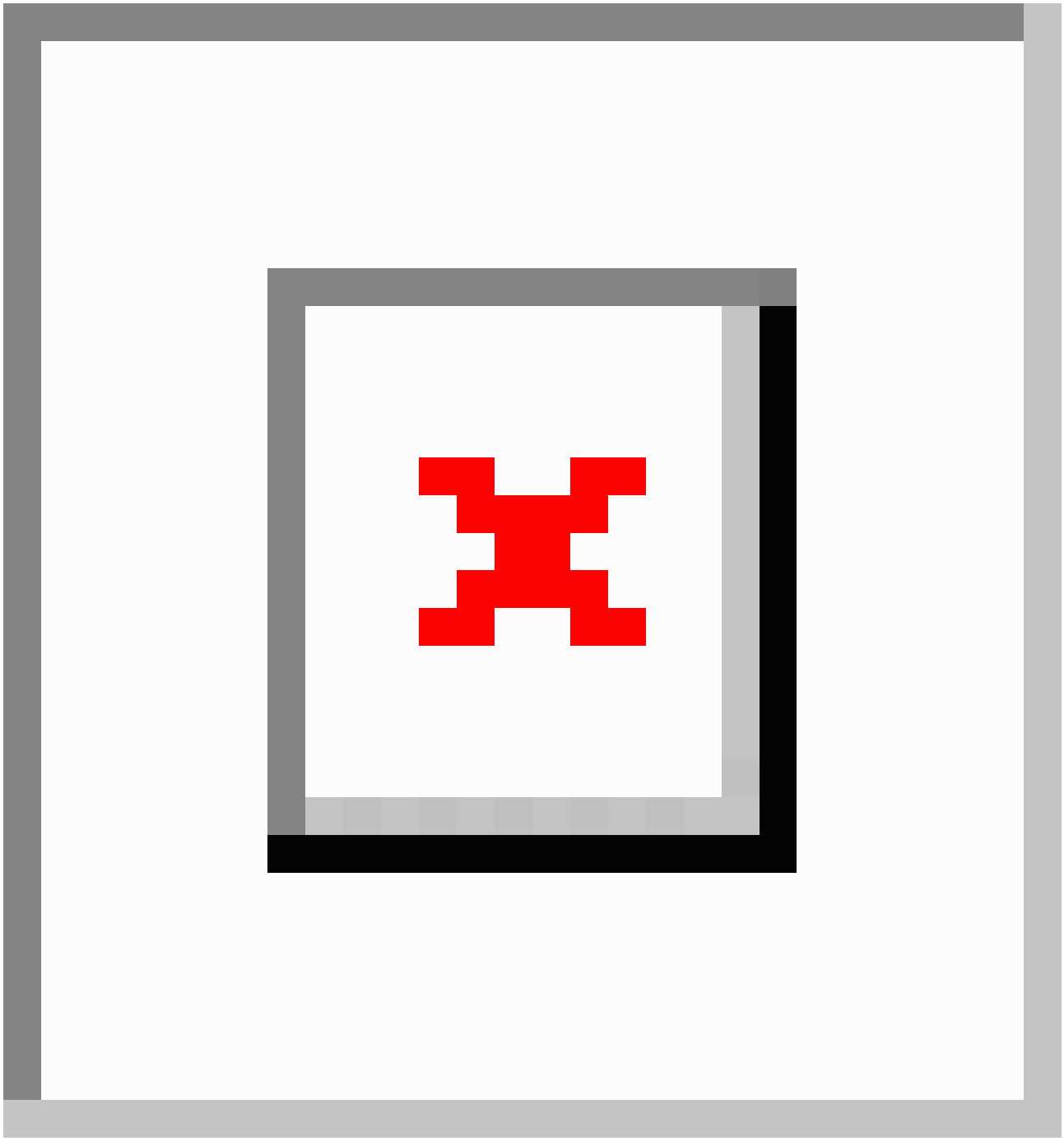


Figure 3 shows the relationship between breast cancer treatment and adverse reactions. Pharmacotherapy is associated with neuropathy, onycholysis, heart neutropenia failure, alopecia, febrile neutropenia, anemia, stomatitis, leukopenia, thrombocytopenia, premature menopause, and gastrointestinal dysfunction. Additionally, multiple nodes are connected, forming multiple pathways, such as pharmacotherapy-febrile neutropenia-adjuvant chemotherapy and pharmacotherapy-leukopenia-breast cancer therapeutic procedure-osteoporosis.

Figure 3. Relationship between breast cancer treatment and adverse reactions.

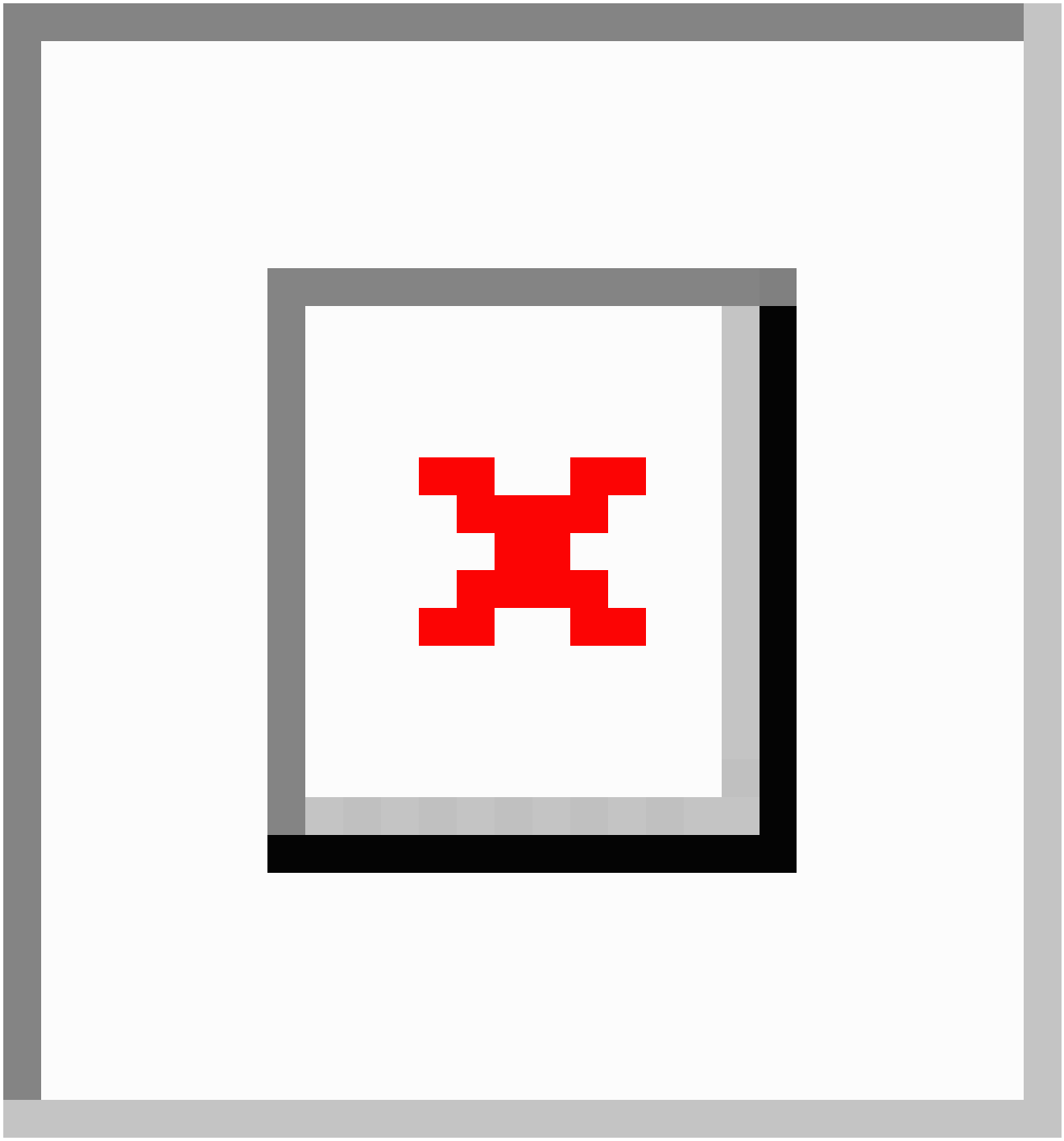
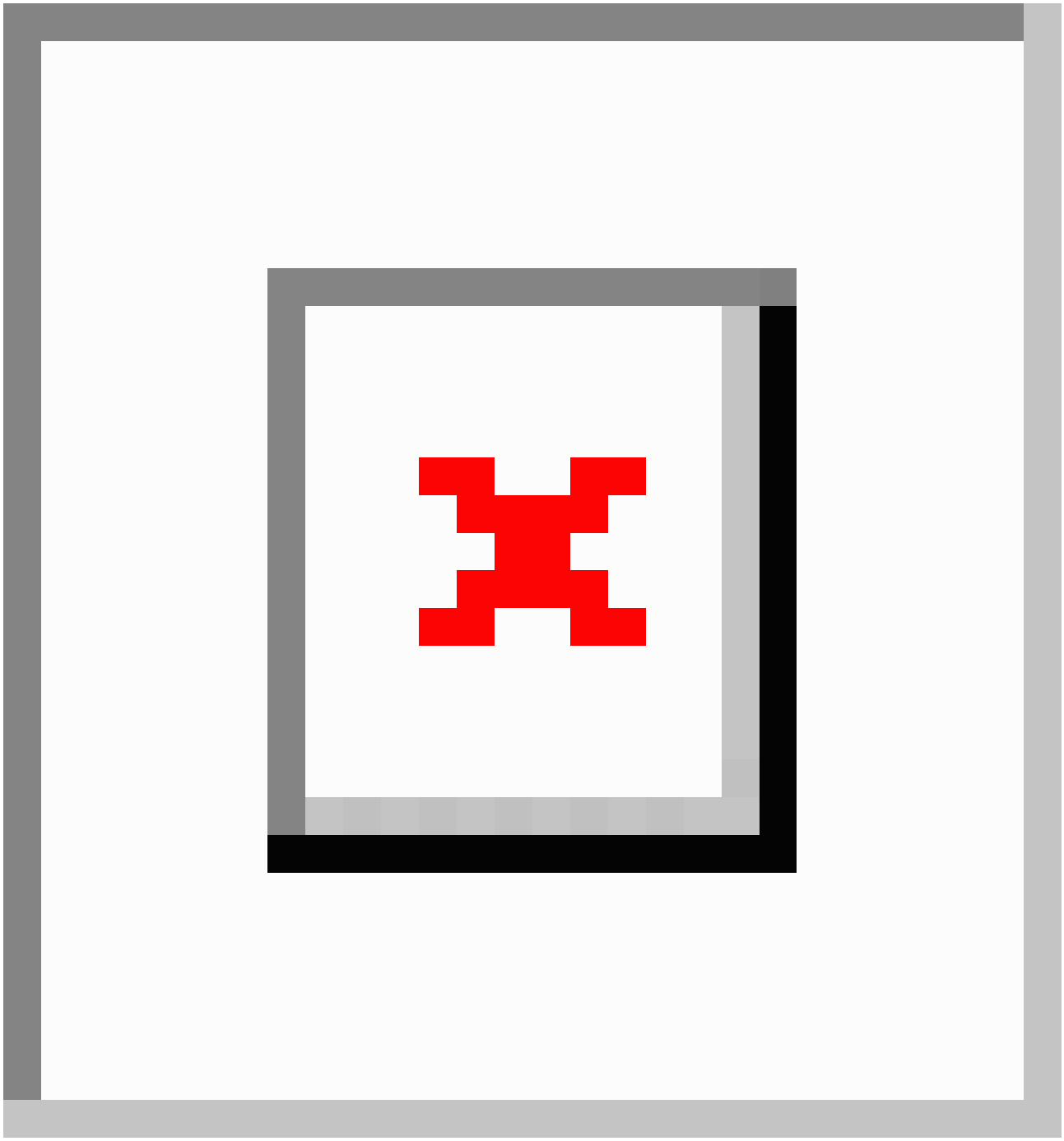


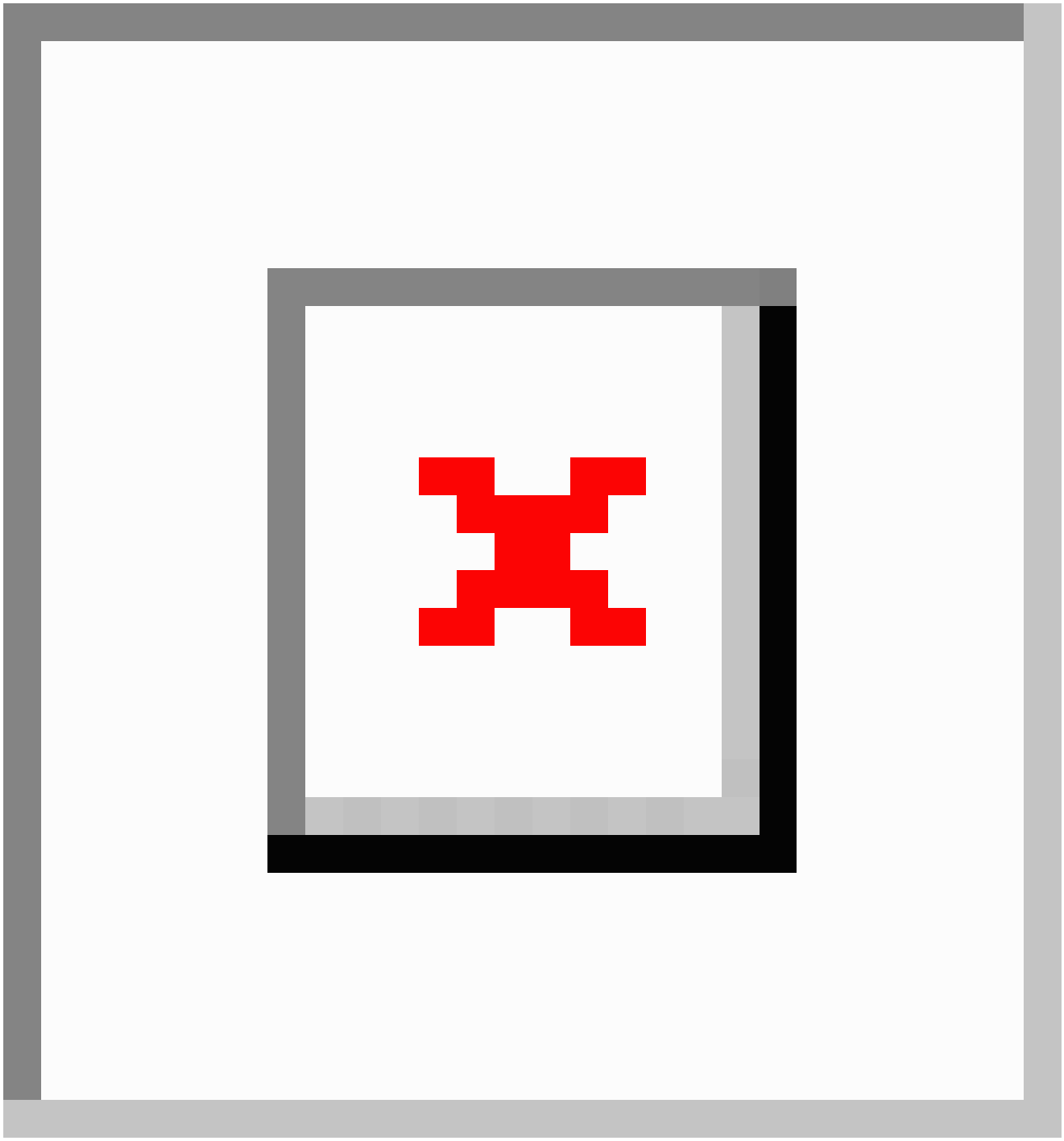
Figure 4 shows the relationship between adverse events after breast cancer treatment and preventive measures. Peripheral neuropathy is associated with cryotherapy, low-level laser therapy, compression procedure, acupuncture procedure, pharmacotherapy, and massage. Lymphedema is associated with resistance education, axillary lymph node dissection, physical therapy, excision of axillary lymph nodes group, and drainage of lymphatics. Early radiation dermatitis is associated with topical administration and bleomycin, cisplatin, or methotrexate protocol. In addition, there are some adverse reactions with relatively few treatment measures, such as stomatitis-diet, alopecia-scalp cooling.

Figure 4. Relationship between adverse reactions after breast cancer treatment and preventive measures.



We performed a relationship visualization to gain a better understanding of the association between types of breast cancer, treatments, drugs, and genes. Figure 5 intuitively reflects the high frequency of malignant neoplasm of the breast, pharmacotherapy, and trastuzumab. In addition, breast malignant tumors are associated with multiple genes, such as the phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA*) gene, platelet-derived growth factor receptor beta (*PDGFRB*) gene, phosphatase and tensin homolog (*PTEN*) gene, and erb-B2 receptor tyrosine kinase 2 (*ERBB2*) gene.

Figure 5. Relationship between types of breast cancer, treatments, drugs, and genes.



Discussion

Principal Findings

The knowledge graphs constructed in this study help researchers understand the research hot spots in breast cancer over the past 5 years. The complex network involving treatment methods, drugs, adverse reactions, preventive measures, and genes in breast cancer can assist clinicians in making decisions that comprehensively consider multiple aspects, ultimately aiding in decisions that are the most beneficial to patients. Additionally, the knowledge graph allows for personalized considerations based on specific genes for individualized patients.

This study found that from 2018 to 2022, breast malignancies appeared most frequently in the literature and were the primary concern for researchers. Research interest in triple-negative breast neoplasms is higher than in other subtypes. This phenomenon may be due to the higher risk of recurrence and poor prognosis in patients with early-stage triple-negative breast neoplasms [10], making it a subject of greater concern to clinicians and researchers. Among treatment modalities, pharmacotherapy receives the highest attention. Pharmacotherapy for breast cancer primarily involves chemotherapy, endocrine therapy, and targeted therapy [27]. Compared to traditional surgery and radiotherapy, pharmacotherapy can more precisely intervene in the growth and division of cancer cells by targeting specific molecules or

cellular structures, which reduces damage to normal cells and allows for the formulation of personalized treatment plans based on the patient's genotype and molecular characteristics [28]. Medications circulating through the bloodstream can also act on cancer cells throughout the body, preventing cancer cell metastasis. These advantages of pharmacotherapy may be related to the heightened emphasis on pharmacotherapy over the past 5 years. Trastuzumab receives the highest attention in breast cancer pharmacotherapy; it is a specific cancer-targeting medication used in the treatment of cancers characterized by elevated levels of HER2 protein [29].

Pharmacotherapy is associated with various adverse reactions, including neutropenia, neuropathy, onycholysis, heart failure, alopecia, and febrile neutropenia. Among these adverse reactions, peripheral neuropathy and lymphedema have the most corresponding preventive and treatment measures, with lymphedema being a common complication after surgery [30]. However, there is limited research on how to prevent and treat the potential adverse reactions of pharmacotherapy, and further studies are needed. Various adverse effects of breast cancer treatment may reduce patients' adherence to treatment. Therefore, when clinicians choose different treatments and drugs, they should pay close attention to their potential adverse reactions and how to prevent or mitigate them.

In existing knowledge graphs related to breast cancer, one study from China constructed a knowledge graph using electronic medical records, clinical guidelines, and expert opinions, primarily focusing on breast cancer diagnosis [18]. Another study by Chinese scholars also used data from various sources, including clinical guidelines, medical encyclopedias, and electronic medical records, to construct a knowledge graph primarily applied to medical knowledge question-answering and medical record retrieval [19]. These studies used data from multiple sources, including structured, unstructured, and semistructured data. Data extraction and accuracy face challenges. Therefore, they used neural network models for training and calculated a series of metrics to ensure data accuracy. For instance, they utilized BERT + Bi-LSTM+ CRF for textual data to achieve named entity recognition. In this study, SemMedDB was used as the data source, and the database was constructed by extracting semantic information from PubMed using SemRep, which demonstrated good performance in a biomedical text [21].

In summary, the knowledge graph constructed in this study for breast cancer treatment and prevention encompasses information on different stages, subtypes of breast cancer, treatment modalities, medications, adverse reactions, and preventive measures. This knowledge forms a complex network, providing clinical practitioners with a comprehensive and referenced knowledge base. We recommend that clinical practitioners apply our research findings in several aspects. First, clinicians can gain insights into the current state of breast cancer treatment and prevention research through our study. Additionally, there is a relative lack of preventive measures and strategies for mitigating postoperative and postmedication adverse reactions compared to breast cancer treatment, and more efforts are needed in these areas. Furthermore, our research can assist clinicians in making comprehensive decisions. For instance, when selecting a treatment approach for patients, the knowledge graph facilitates linking to available medications, associated adverse reactions, and measures to mitigate or prevent adverse effects.

Our research still has several limitations. First, SemRep, as a natural language processing program based on the UMLS, still exhibits shortcomings. Despite the extensive coverage and scale of the UMLS Metathesaurus, it has a relatively limited ability to recognize entities. There are still areas for improvement in processing natural language texts [20]. Second, clinical researchers often prefer causal relationships rather than pure correlations; however, our study can only reveal the connections between pieces of information and cannot determine the magnitude and direction of their effects. Third, with the release of new literature, the knowledge graph also needs to be updated promptly, increasing the burden on researchers. Future improvements should focus on automating the mining of literature data to ensure timely updates to the knowledge graph for breast cancer prevention and treatment, thereby alleviating the burden on researchers.

Conclusions

This study successfully constructed a knowledge graph for breast cancer prevention and treatment by integrating relevant literature from the past 5 years and conducting knowledge discovery. Through this knowledge graph, researchers can learn about breast cancer treatment methods, medications, and adverse reactions to preventive treatments and gain insights into the relationships between different pieces of knowledge.

Acknowledgments

The authors would like to thank Feng Xixi, associate chief physician and member of the Chronic Disease Special Committee of the Chengdu City Preventive Medicine Association, for her suggestions at the initial stage of the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table depicting the semantic relationship and semantic schema of breast cancer.
[DOCX File, 19 KB - medinform_v12i1e52210_app1.docx]

Multimedia Appendix 2
Different subtypes and stages of breast cancer.
[PNG File, 158 KB - medinform_v12i1e52210_app2.png]

Multimedia Appendix 3
Treatments of breast cancer.
[PNG File, 214 KB - medinform_v12i1e52210_app3.png]

Multimedia Appendix 4
Drugs for breast cancer.
[PNG File, 160 KB - medinform_v12i1e52210_app4.png]

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018 Nov;68(6):394-424. [doi: 10.3322/caac.21492] [Medline: 30207593]
2. Xiao Y, Xia J, Li L, et al. Associations between dietary patterns and the risk of breast cancer: a systematic review and meta-analysis of observational studies. Breast Cancer Res 2019 Jan 29;21(1):16. [doi: 10.1186/s13058-019-1096-1] [Medline: 30696460]
3. Sung H, Ferlay J, Siegel RL, 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. [doi: 10.3322/caac.21660] [Medline: 33538338]
4. Thakur P, Seam RK, Gupta MK, Gupta M, Sharma M, Fotedar V. Breast cancer risk factor evaluation in a Western Himalayan state: a case-control study and comparison with the Western World. South Asian J Cancer 2017;6(3):106-109. [doi: 10.4103/sajc.sajc_157_16] [Medline: 28975116]
5. Badr LK, Bourdeanu L, Alatrash M, Bekarian G. Breast cancer risk factors: a cross- cultural comparison between the west and the east. Asian Pac J Cancer Prev 2018 Aug 24;19(8):2109-2116. [doi: 10.22034/APJCP.2018.19.8.2109] [Medline: 30139209]
6. Zhang X, Dong XP, Guan YZ, Me R, Guo DL, He YT, et al. Research progress on epidemiological trend and risk factors of female breast cancer. Cancer Res Prev Treat 2021;48(1):87-92.
7. Tan MM, Ho WK, Yoon SY, et al. A case-control study of breast cancer risk factors in 7,663 women in Malaysia. PLoS One 2018;13(9):e0203469. [doi: 10.1371/journal.pone.0203469] [Medline: 30216346]
8. Britt KL, Cuzick J, Phillips KA. Key steps for effective breast cancer prevention. Nat Rev Cancer 2020 Aug;20(8):417-436. [doi: 10.1038/s41568-020-0266-x] [Medline: 32528185]
9. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. In: Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004: Association for Computational Linguistics; 2004:76-83. [doi: 10.5555/1596431.1596442]
10. For the progress of adjuvant treatment of triple-negative breast cancer, just look at these 8 key clinical studies! [Article in Chinese]. Sohu. 2021 Dec 14. URL: https://www.sohu.com/a/508222106_121118854 [accessed 2023-06-25]
11. Feng B, Gao J. AnthraxKP: a knowledge graph-based, anthrax knowledge portal mined from biomedical literature. Database (Oxford) 2022 Jun 2;2022:baac037. [doi: 10.1093/database/baac037] [Medline: 35653350]
12. Feng F, Tang F, Gao Y, et al. GenomicKB: a knowledge graph for the human genome. Nucleic Acids Res 2023 Jan 6;51(D1):D950-D956. [doi: 10.1093/nar/gkac957] [Medline: 36318240]
13. James T, Hennig H. Knowledge graphs and their applications in drug discovery. Methods Mol Biol 2024;2716:203-221. [doi: 10.1007/978-1-0716-3449-3_9] [Medline: 37702941]
14. Lyu K, Tian Y, Shang Y, et al. Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy. J Biomed Inform 2023 Mar;139:104298. [doi: 10.1016/j.jbi.2023.104298] [Medline: 36731730]
15. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017 Jan 4;45(D1):D833-D839. [doi: 10.1093/nar/gkw943] [Medline: 27924018]
16. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018 Jan 4;46(D1):D1074-D1082. [doi: 10.1093/nar/gkx1037] [Medline: 29126136]
17. ClinVar. National Library of Medicine. URL: https://www.ncbi.nlm.nih.gov/clinvar [accessed 2023-11-18]
18. Li X, Sun S, Tang T, et al. Construction of a knowledge graph for breast cancer diagnosis based on Chinese electronic medical records: development and usability study. BMC Med Inform Decis Mak 2023 Oct 10;23(1):210. [doi: 10.1186/s12911-023-02322-0] [Medline: 37817193]
19. An B. Construction and application of Chinese breast cancer knowledge graph based on multi-source heterogeneous data. Math Biosci Eng 2023 Feb 6;20(4):6776-6799. [doi: 10.3934/mbe.2023292] [Medline: 37161128]

20. Li XY, Li JL, Li ZY. Integrated medical language system and its application in knowledge discovery. Digital Library Forum 2019;9:24-29.
21. Kilicoglu H, Rosemblat G, Fiszman M, Shin D. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics 2020 May 14;21(1):188. [doi: 10.1186/s12859-020-3517-7] [Medline: 32410573]
22. Access to SemRep/SemMedDB/SKR resources. National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR.html [accessed 2023-11-18]
23. McKinney W. Pandas: a foundational Python library for data analysis and statistics. In: Python for High Performance and Scientific Computing: Deutsches Zentrum für Luft-und Raumfahrt; 2010:293-296.
24. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9(3):90-95. [doi: 10.1109/MCSE.2007.55]
25. WordCloud for Python documentation. Andreas C. Müller - Machine Learning Scientist. URL: https://amueller.github.io/word_cloud/ [accessed 2023-12-25]
26. Hagberg A, Swart PJ, Schult DA. Exploring Network Structure, Dynamics, and Function Using NetworkX: Los Alamos National Lab (LANL); 2008.
27. The difference between breast cancer radiotherapy, targeted therapy and chemotherapy! [Article in Chinese]. Sohu. 2018 Dec 7. URL: https://www.sohu.com/a/280208482_790163 [accessed 2023-11-18]
28. Nagini S. Breast cancer: current molecular therapeutic targets and new players. Anticancer Agents Med Chem 2017;17(2):152-163. [doi: 10.2174/1871520616666160502122724] [Medline: 27137076]
29. Trastuzumab. Cancer Research UK. URL: https://www.cancerresearchuk.org/about-cancer/treatment/drugs/trastuzumab [accessed 2023-11-18]
30. Bernas M, Thiadens SRJ, Smoot B, Armer JM, Stewart P, Granzow J. Lymphedema following cancer therapy: overview and options. Clin Exp Metastasis 2018 Aug;35(5-6):547-551. [doi: 10.1007/s10585-018-9899-5] [Medline: 29774452]

Abbreviations

ERBB2: erb-B2 receptor tyrosine kinase 2
HER2: human epidermal growth factor receptor 2
PDGFRB: platelet-derived growth factor receptor beta
PIK3CA: phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PTEN: phosphatase and tensin homolog
SemMedDB: Semantic MEDLINE Database
SPO: Subject-Predicate-Object
UMLS: Unified Medical Language System

Edited by A Benis; submitted 26.08.23; peer-reviewed by C Gaudet-Blavignac, S Yang, Y Chu; revised version received 02.01.24; accepted 06.01.24; published 22.02.24.

Please cite as:
Jin S, Liang H, Zhang W, Li H
Knowledge Graph for Breast Cancer Prevention and Treatment: Literature-Based Data Analysis Study
JMIR Med Inform 2024;12:e52210
URL: https://medinform.jmir.org/2024/1/e52210
doi:10.2196/52210

Evaluation of SNOMED CT Grouper Accuracy and Coverage in Organizing the Electronic Health Record Problem List by Clinical System: Observational Study

Rashaud Senior[1], MMCi, MD; Timothy Tsai[2], MMCi, DO; William Ratliff[3], MBA; Lisa Nadler[1], MD; Suresh Balu[3], MS, MBA; Elizabeth Malcolm[4], MSHS, MD; Eugenia McPeek Hinz[1], MS, MD

1

2

3

4

Corresponding Author:
Rashaud Senior, MMCi, MD

Abstract

Background: The problem list (PL) is a repository of diagnoses for patients' medical conditions and health-related issues. Unfortunately, over time, our PLs have become overloaded with duplications, conflicting entries, and no-longer-valid diagnoses. The lack of a standardized structure for review adds to the challenges of clinical use. Previously, our default electronic health record (EHR) organized the PL primarily via alphabetization, with other options available, for example, organization by clinical systems or priority settings. The system's PL was built with limited groupers, resulting in many diagnoses that were inconsistent with the expected clinical systems or not associated with any clinical systems at all. As a consequence of these limited EHR configuration options, our PL organization has poorly supported clinical use over time, particularly as the number of diagnoses on the PL has increased.

Objective: We aimed to measure the accuracy of sorting PL diagnoses into PL system groupers based on Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) concept groupers implemented in our EHR.

Methods: We transformed and developed 21 system- or condition-based groupers, using 1211 SNOMED CT hierarchal concepts refined with Boolean logic, to reorganize the PL in our EHR. To evaluate the clinical utility of our new groupers, we extracted all diagnoses on the PLs from a convenience sample of 50 patients with 3 or more encounters in the previous year. To provide a spectrum of clinical diagnoses, we included patients from all ages and divided them by sex in a deidentified format. Two physicians independently determined whether each diagnosis was correctly attributed to the expected clinical system grouper. Discrepancies were discussed, and if no consensus was reached, they were adjudicated by a third physician. Descriptive statistics and Cohen κ statistics for interrater reliability were calculated.

Results: Our 50-patient sample had a total of 869 diagnoses (range 4-59; median 12, IQR 9-24). The reviewers initially agreed on 821 system attributions. Of the remaining 48 items, 16 required adjudication with the tie-breaking third physician. The calculated κ statistic was 0.7. The PL groupers appropriately associated diagnoses to the expected clinical system with a sensitivity of 97.6%, a specificity of 58.7%, a positive predictive value of 96.8%, and an F_1-score of 0.972.

Conclusions: We found that PL organization by clinical specialty or condition using SNOMED CT concept groupers accurately reflects clinical systems. Our system groupers were subsequently adopted by our vendor EHR in their foundation system for PL organization.

(JMIR Med Inform 2024;12:e51274) doi:10.2196/51274

KEYWORDS

electronic health record; problem List; problem list organization; problem list management; SNOMED CT; SNOMED CT Groupers; Systematized Nomenclature of Medicine; clinical term; ICD-10; International Classification of Diseases

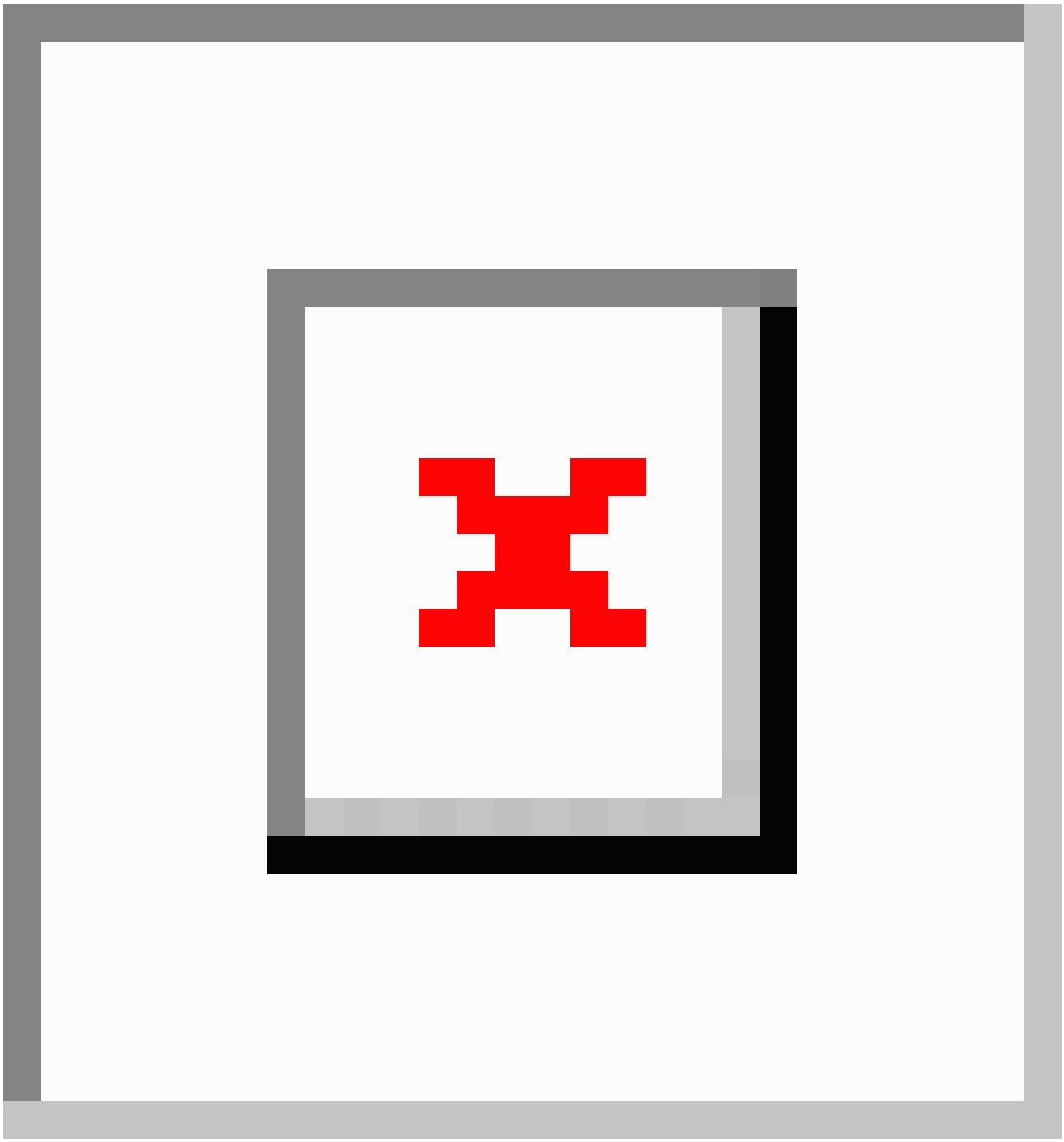
Introduction

The electronic health record (EHR) problem list (PL) is a dynamic repository of a patient's current and historical conditions as well as other health-related issues. As such, it supports communication across a wide range of potential caregivers and clinical environments. An accurate PL serves as a foundation for clinical care and population health management, with multiple derivative secondary processes, including phenotype extraction and disease prediction.

Understanding the history of the PL helps to illustrate why this construct has become the default format for summarizing patients' clinical history. In the 1960s, Lawrence L Weed, MD, proposed the concepts of the problem-oriented medical record; the PL; and the Subjective, Objective, Assessment, and Plan (SOAP) notes for documentation [1]. The idea was to colocate clinical problems with clinical results to focus on systematically addressing all of a patient's diagnoses [2]. Although the SOAP note became the standard format for clinical notes, the PL has encountered more inconsistent use, struggling with problems of inaccuracy, missing diagnoses, not being updated, and bloating [3]. In 2009, the HITECH (Health Information Technology Economic and Clinical Health) Act codified the requirement for an up-to-date PL for meaningful use [4]. Until recently, our vendor EHR had relied on relatively ineffective organization strategies for the PL.

With no one owner, the PLs have become disorganized and cluttered with duplications, conflicting entries, and no-longer-valid diagnoses that contribute to information overload and bloat, obscuring the patient's clinical picture [5]. In its former state, our EHR PL was organized primarily alphabetically, with other options based on primary specialty or priority, all of which have limited clinical utility, especially as the number of diagnoses on the PL increases (Figure 1) . For example, for one patient, we found active diagnoses of lung nodule (Respiratory System), then lung cancer (Oncology System), and then lung cancer with brain metastases (Oncology System). These diagnoses were all related to the same problem but were added sequentially with previous diagnoses that were no longer clinically relevant and were not removed.

Figure 1. Appearance of a problem list before and after grouping algorithm application. Items were reorganized into 21 system groupers using Boolean logic with the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) codes; they were then translated into the *International Classification of Diseases, Tenth Version* (*ICD-10*) codes. Groupers were based on a combination of traditional medical specialty categories, clinically relevant care coordination, and procedure-based groupings, some of which were themselves combined due to overlapping diagnostic coverage. The final order of the problem list items was determined by Epic System's base hierarchy. CMS: The Centers for Medicare and Medicaid Services; HCC: Hierarchical Condition Category; HHS: US Department of Health and Human Services; FEN/GI: Fluids, Electrolytes, Nutrition/Gastrointestinal; GFR=Glomerular Filtration Rate.



There are several major terminology standards that capture patient diagnoses, symptoms, and other health-related conditions, two of which are the *International Classification of Diseases* (*ICD*) [6] and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [7]. The World Health Organization maintains *ICD* codes, which are designed to classify diseases, conditions, and other health-related issues [8,9]. Organized into 21 chapters, they use an alphanumeric classification format to identify diseases, injuries, and factors influencing health. The United States uses an additional Clinical Modifier for further specificity [10]. The *ICD* codes are used in various clinical and nonclinical settings, including disease description, treatment selection, billing, and research applications [11]. There are 78,044 total codes in the 2024 code set, according to the Centers for Medicare and Medicaid Services [12].

Currently managed by Systematized Nomenclature of Medicine (SNOMED) International (previously known as the International

Health Terminology Standards Development Organization [IHTSDO]), SNOMED CT was designed to be a US standard for health information exchange. It functions as a highly granular ontology used to describe clinical observations and findings [13]. SNOMED CT uses a polyhierarchal (parent-child) format organized around a general root concept (eg, a clinical finding, procedure, situation with explicit context, or event) with increased granularity achieved by differentiating more specific descriptions of that root concept. This process allows for the representation of specific clinical content in a machine-readable format [14,15]. Updated monthly, the total number of concepts is 512,087 (as of April 1, 2024 [16]) and continues to increase over time.

Though required for billing, the *ICD, Tenth Revision* (*ICD-10)* terminology is not used directly in clinical care, as many code names are not consistent with clinical vernacular. For example, code Z91.038 "Allergy status to unspecified drugs, medicaments and biological substances" is not as intuitive as "Allergy to insect stings." For this example, our third-party vendor, Intelligent Medical Objects (IMO), transforms the *ICD-10* codes into clinically relevant human-readable concepts. IMO additionally maps at least 1 SNOMED CT concept for each diagnosis, attached as metadata, upon which the PL groupers can be organized. Although SNOMED CT concepts are mapped to *ICD-10* codes [17,18], as with most ontologies, there are gaps in clinical concept coverage.

PL organization and cleanup is challenging for many reasons, including there being no single owner of a patient's PL [19] and its maintenance and cleanup being secondary to other direct clinical care priorities. The tools in the EHR for cleanup are limited to a single patient and do not allow for automated processing or opportunities to categorize or define the state of the PL or large-scale maintenance at the population level. Multiple interventions, including reconfiguration of the EHR PL and re-education, have been met with limited success [20]. Despite these attempts, PL bloat and inaccuracy are widely recognized as issues affecting clinical care and secondary downstream uses of the data [3]. We have come to recognize that curating a clinically relevant and updated PL is a difficult challenge; our primary option for improving its organization was to extend and improve SNOMED CT groupers.

In this paper, we present a PL reorganization developed around clinical specialty groupings using SNOMED CT codes and Boolean logic. We describe the evaluation of the new PL groupers for clinical accuracy and efficiency using a convenience set of patients and their diagnoses. This system allows for future characterization of the PLs at the patient and population levels; it also provides potential for automated cleanup options in the future.

Methods

SNOMED CT Grouper Development and Evaluation

Author EMH extended and extensively modified 19 previously defined groupers initially developed by Heidi Twedt, MD, and added 2 newly defined system- or condition-based groupers, one for pediatric and one for transplant-specific conditions (Table 1). System groupers included traditional medical specialty categories as well as clinically relevant care coordination and procedure-based groupings. Some specialties were combined due to overlapping diagnosis domains (eg, "Respiratory and Allergy" and "Orthopedic and Musculoskeletal" domains). The primary focus was for the system grouper diagnoses to be organized around clinical use. For example, "acute myocardial infarction" and "venous thromboembolism" were sorted into the "Cardiovascular and Peripheral Vascular" grouper, while addiction issues, such as "alcohol use disorder," were sorted into the "Behavioral Health" grouper.

Table . List of system groupers with example diagnoses.

Condition or specialty grouper	Example diagnosis	Notable deviations
1. Care Coordination	Physical deconditioning, food insecurity, risk for falls	Includes health-related social needs
2. Oncology	Malignancies and radiation therapy diagnoses	Excludes dermatology cancers and includes treatment complications
3. Cardiovascular and Peripheral Vascular	Atrial fibrillation and deep vein thrombosis	Excludes cerebral vascular diagnoses
4. Respiratory and Allergy	Asthma and peanut allergy	—[a]
5. Endocrine	Diabetes mellitus, gout, and hypothyroidism	—
6. Behavioral Health	Schizophrenia and opioid use disorder	—
7. Transplant	Living-related kidney transplant and graft versus host disease	Includes transplant complications
8. Infectious Disease, Immune, or Lymphatic	Pneumonia and immune deficiencies	—
9. Blood	Anemia	—
10. Neurology or Sleep	Seizure and sleep disorders	Excludes chronic pain
11. Ears, Nose, Throat (ENT)	Nasal polyps, cleft palate, and hearing loss	—
12. Fluids, Electrolytes, Nutrition, and Gastrointestinal	Hyponatremia and Crohns disease	—
13. Obstetrics and Gynecology	Ovarian cysts; hemolysis, elevated liver enzymes, low platelet count (HELLP) syndrome; and dense breast tissue	Female-specific diagnoses
14. Genitourinary and Nephrology	Ureteral calculus and prostatitis	Includes male-specific genitourinary issues
15. Dermatology	Atopic dermatitis and melanoma	Includes all dermatology-specific cancers (eg, squamous or basal cell carcinoma)
16. Rheumatology	Rheumatoid arthritis	—
17. Orthopedic and Musculoskeletal	Hip fracture	—
18. Ophthalmology or Eye	Uveitis	Includes complications of eye from other diseases
19. Genetics	Trisomy 21	Includes all nonspecific system genomic issues
20. Pediatrics	28-week prematurity	Includes developmental disorders
21. Surgery, Trauma, Wound, and Pain	Gunshot wound and complex regional pain syndrome	—
22. Other	Edema and medication management	Includes any diagnosis that does not fit into another grouper

[a]Not applicable.

Due to its polyhierarchal framework, all child-related SNOMED CT concepts include all related downstream concepts, unless excluded by the Boolean logic. In this format, fewer SNOMED CT concepts can represent many derivative *ICD-10* codes more comprehensively than could be achieved by directly curating *ICD-10* codes. For example, 167 SNOMED CT concepts within the Neurology grouper were mapped to 9243 IMO *ICD-10* diagnoses. Our default EHR PLs were reorganized according to this system-based methodology in the order presented in Table 1.

Using our EHR vendor's built-in tools for grouper build, author EMH iteratively refined the groupers to be consistent with clinical systems using 1211 SNOMED CT concepts. System groupers included the highest parent concept that was appropriate with logic to exclude child-related SNOMED CT concepts not clinically appropriate for a system. For example, squamous cell carcinoma and skin cancers in general are managed clinically by dermatology. In the build for the Oncology grouper, therefore, all dermatologic cancers were excluded and instead added to the Dermatology grouper. As another example, our Cardiovascular grouper includes peripheral vascular diseases like "deep venous thrombosis" but excludes cerebral vascular concepts. This allows diagnoses like "cerebral avascular malformation" to be presented within our Neurology grouper. Both examples highlight the focus of this grouper organization to support clinical specialty coordination of diagnoses.

Multisystem disorders were grouped according to the specialty that would typically manage each disease entity. For example, systemic lupus erythematosus was grouped under "Rheumatology." If a diagnosis's SNOMED CT concept was too broad to be captured by one of the 21 groupers, it defaulted

into the "Other" category. For example, "edema" is a clinical finding that can be reasonably attributed to multiple diseases. As such, it does not have a specific condition or specialty and instead falls into the "Other" category.

To evaluate the effectiveness of specialty sorting, we used a convenience sample of 50 patients randomly identified in January 2022. These patients had at least 3 encounters in the previous year and were selected across all age groups, ranging from newborn to geriatric patients, with an equal ratio of sexes (Table 2). The encounter criteria ensured that identified patients had multiple recent opportunities to have their PLs updated. The PLs for these patients were extracted through screen capture software by author EMH to develop a cohort with no patient identifiers. Standard EHR PL functionality included system grouper name, time frame since the problem was added to the PL, and a limited free-text overview if included with the entry. These study PL entries were reconfigured into a study document with labels indicating sequential patient number, patient age, and patient sex.

Two of the authors, both family medicine physicians (TT and RS), independently examined each patient's PL to determine the clinical accuracy of system groupings for all diagnoses (Table 3). For any items whose system attribution they questioned, the reviewers identified the SNOMED CT code attached to the *ICD-10* code. Diagnoses that were deemed correctly grouped into the appropriate system grouper were considered true positives, while those that were incorrectly grouped were considered false positives. All diagnoses in the dropout "Other" category were examined by their associated SNOMED CT code for options for attribution to a defined system grouper. A diagnosis for which the SNOMED CT code was too vague or not specific enough to be grouped was considered a true negative. Any diagnosis that had a SNOMED code that could have been placed in a relevant system grouper but was not was considered a false negative.

Table . Patient demographics and baseline descriptive statistics. A total of 50 patients, subdivided by age and sex, with descriptive statistics, were reported for each age range.

Age ranges (years)	Gender			Problems			
	Total, n (%)	Male, n (%)	Female, n (%)	Total, n	Mean	Median	Min-Max
<1	6 (12)	3 (50)	3 (50)	72	12.0	10	6-20
1-17	7 (14)	3 (42.9)	4 (57.1)	157	22.4	26	4-35
18-64	24 (48)	11 (45.8)	13 (54.2)	342	14.3	10	4-43
≥65	13 (26)	8 (61.5)	5 (38.5)	298	22.9	21	4-59
All ages	50 (100)	25 (50)	25 (50)	869	17.4	12	4-59

Table . Description of metrics used to determine the effectiveness of automated system grouping. Two reviewers examined individual problem list items and their assigned grouping, placing each into a category.

Assessment category	Definition	Example
True positive (correct system association)	Diagnosis falls into the right disease system—the SNOMED[a] grouper is specific and attributable.	"Community-acquired pneumonia" in the Infectious Disease system
False positive (incorrect system association)	Diagnosis falls in the wrong system grouper.	"Diaphragmatic stimulation by cardiac pacemaker" grouped under "Central Hypoventilation Syndrome".
True negative (Other—correct system association)	The SNOMED grouper associated with a diagnosis is not specific enough to be in anything but the Other category.	"Anticoagulated" placed with the SNOMED grouper "Drug therapy finding". This is not specific enough to be attributed to just anticoagulation status.
False negative (Other—incorrect system association)	Diagnosis belongs to a specified system grouper but falls into the Other category due to logic deficits in the grouper.	"Genetic disorder" falling into the "Other" category until the VCG Grouper is corrected.

[a]SNOMED: Systematized Nomenclature of Medicine.

Each diagnosis was independently categorized according to the scheme in Table 3; then the reviewers compared their determinations. A third independent clinician (author LN) served as a tie-breaker for those PL items for which an agreement was not reached. We calculated descriptive statistics to summarize the volume of diagnoses for the 50 test patients and performance metrics to assess the accuracy and validity of the groupers. The correlation coefficient (κ statistic) was calculated for the degree of agreement between reviewers and for SNOMED CT grouper attributions.

This work was performed using Epic Systems (version May 2021; Verona, WI) initially deployed with ambulatory applications in July 2012 and inpatient applications in June 2013 within the Duke University Health System.

Ethical Considerations

All patient data were anonymized with all demographic identifiers removed except for age. This study was approved by the Duke University Internal Review Board for exempt status (IRB #PRO-00108903).

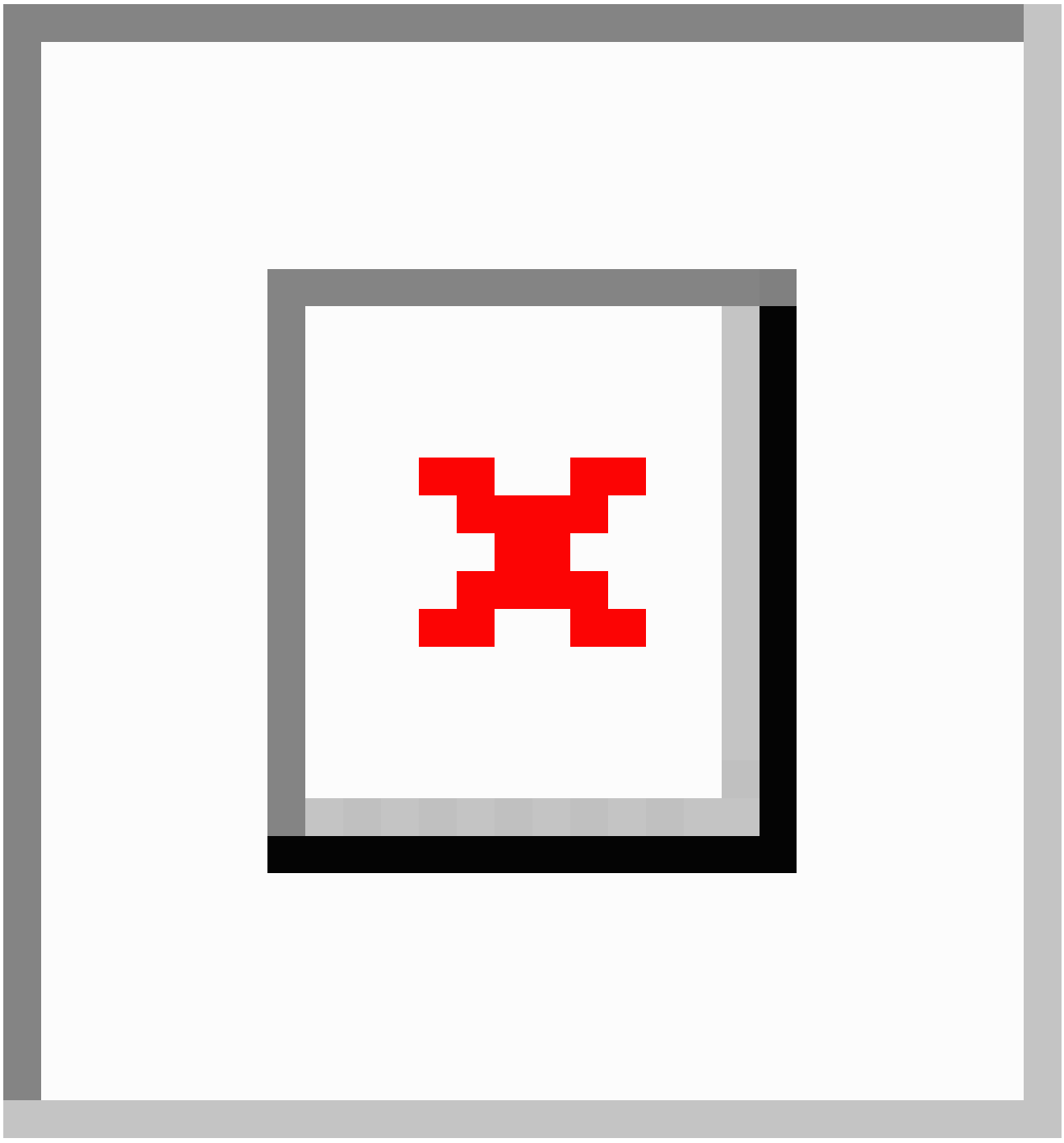
Results

Across the 50 patients, aged 14 days to 93 years, there were a total of 869 (range 4-59) diagnoses identified, with a median of 12 diagnoses per patient. Table 2 includes the breakdown of the volume of PL entries across age and sex.

After their independent evaluations of the PLs, the reviewers initially agreed on 821 (94.4%) of the 869 total problems (Cohen κ coefficient of 0.7, indicating moderate agreement [21]). Of the remaining 48 diagnoses, they subsequently agreed on 32 for a revised agreement rate of 98.2%. The remaining 16 were adjudicated by author LN for attribution.

Based on the definitions presented in Table 3, Figure 2 describes our results. Our final attribution evaluation found that the diagnoses were correctly attributed to a system grouper (ie, sensitivity) in 97.6% of cases, and the nonspecific diagnoses were correctly placed in the "Other" category (ie, specificity) in 58.7% of cases. The positive predictive value, or the correct grouper accuracy rate, was 96.8%. We found 37 (4.3%) true negatives, representing concepts without a SNOMED CT code or diagnoses too general to be attributed to a clinical system. The calculated F_1-Score was 0.972.

Figure 2. Two clinicians' review of problem list sorting algorithm. FN: false negative; FP: false positive; PPV: positive predictive value; NPV: negative predictive value; TN: true negative; TP: true positive; Sn: sensitivity; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms; Sp: specificity.



Discussion

Overview

The PL is the repository of medical diagnoses intended to reflect the patient's clinical conditions. Without groupings reflective of the larger specialty formats of clinical care, the PL can become overloaded and difficult to use as a tool to communicate a patient's clinical status across encounters. We developed 21 SNOMED CT groupers for system concepts to standardize the organization of our EHR PL based on 1211 concepts (Table 1). We chose to evaluate these PL groupers across all ages and sexes to provide a more representative sample of diagnoses across our EHR patient population, recognizing that this is only a subset of the total potential diagnoses. Taking advantage of the hierarchal logic of the SNOMED CT concepts refined with Boolean logic allowed for more than 95% of the diagnoses to be attributed to a system grouper [22,23].

We established the effectiveness of these SNOMED CT groupers in organizing the PL by clinical system related to clinical specialty or condition. We propose that this standardized format for PL organization permits the sharing and reproduction of concepts across other health systems and EHRs.

Comparison to Prior Work

Other groups have used conceptually similar methods with SNOMED CT codes for clinical phenotyping [22]. However, those code sets are typically more narrow in scope for a more specific clinical description. The United Medical Language System Clinical Observations Recordings and Encoding Problem List Subset is meant to "facilitate the use of SNOMED CT as the primary coding terminology for PLs or other summary level clinical documentation" [24]. Compared to these other SNOMED CT code sets, our implementation includes broader clinical coordination groupings (eg, surgical, transplant, care coordination, and infectious disease) that are more reflective of the PL clinical care needs within our institution. Our work here builds upon those efforts and applies them at the system level, which is more accessible for clinical use.

Limitations

We noted some limitations and challenges in using SNOMED CT concepts for this build. Despite ongoing international mapping efforts [17,18], SNOMED CT concepts are not fully representative of all the *ICD-10* codes because of differences in original intended uses [8,13] and baseline granularity [25,26]. For example, the *ICD-10* code "Encounter for pre-transplant evaluation for chronic liver disease" is mapped to the SNOMED CT concept "patient encounter status," as there is no other comparable SNOMED CT coding option. Estimates for the proportions of completely mapped concepts or codes are found in studies reviewing the automation of mapping SNOMED CT and *ICD-10* codes; one study estimated the proportion of complete mappings to *ICD-10-Clinical Modification* (*ICD-10-CM*) at 74% in 2012 [25], and another one estimated the proportion of complete mappings to *ICD-10-Procedure Coding System* (*ICD-10-PCS*) (used to capture inpatient procedures) to be about 86% in 2017 [27].

There were many *ICD-10* diagnoses that were too broad to easily match a SNOMED CT system grouper. "Fatigue" is a good example of an inherently vague constitutional or multisystemic symptom that does not have a clearly identifiable system-level grouper in our schema. For these diagnoses, the "Other" category was used to capture the remaining nonspecific diagnoses. It is important to note that this category is not the same as the *ICD-10* options for "Not Otherwise Specified" (NOS) or "Not Elsewhere Classifiable" (NEC) codes for lesser defined diagnoses. For example, "Pneumonia due to other infectious organisms, NEC" still falls into our "Infectious Disease, Immune, Lymphatic" grouper.

We also note that the mappings are not completely represented across all specialties in terms of the breadth of coverage of concepts. For example, we found more SNOMED CT cardiology-specific concepts and fewer pediatric-specific concepts. These differences may reflect the relative volume of cardiology diagnoses in the general population. The more specific diagnosis of "Encounter for assessment of implantable cardioverter-defibrillator" was mapped to an appropriate SNOMED CT concept and was correctly placed into our cardiovascular system grouper. However, the pediatric diagnosis "Concern about growth" was only mapped to the SNOMED CT code "Finding reported by subject or history provider," which was too broad to be added to the Pediatric grouper only, consequently falling into the "Other" category. Specialties such as pediatrics also require greater levels of specificity for their diagnoses than is always possible with the SNOMED CT concepts currently available.

There were also multiple *ICD-10* codes mapped to the same SNOMED CT code that made attribution to a system grouper challenging. For example, the diagnoses "Diaphragmatic stimulation by pacemaker" and "Disorder of cardiac pacemaker system" mapped to the same SNOMED CT code of "Disorder of cardiac pacemaker system," placing them into the Cardiovascular grouper, although the former would ideally be attributed to the Pulmonary grouper.

As we consider the future challenges of algorithm-based PL sorting, it will be important to investigate the implications of updating ontologies as the World Health Organization has already published the 11th edition of *ICD* (*ICD-11)* with 35 countries now implementing it [28]. We do not suspect that *ICD-11* will replace SNOMED CT as an ontology organization method, as SNOMED CT maintains greater flexibility for clinical use. Health systems are always evolving, and it will be important to consider how such algorithms and their applications will evolve within them.

Conclusions

We leveraged a PL sorting algorithm based on the clinical system–based SNOMED CT groupers to create a standardized PL format in our EHR, reorganizing the diagnoses, symptoms, and medical problems for better clinical utility. We found subjective positive outcomes for our clinical users who reported streamlining their clinical review processes and easier ability to identify similar and duplicate diagnoses. This may be especially helpful for patients with complex issues and many associated diagnoses. A structured PL also enables a shift from patient-level evaluation to potentially population-level assessments and cleanup automation.

As with improvements in the provider experience, automated PL maintenance may also impact researchers leveraging PL diagnoses for machine learning and other similar research. Such possibilities underscore the need for accurate and updated PL diagnoses to achieve and maintain high-fidelity outputs. It will be important to further evaluate methods to automate the maintenance of accurate Pls and best influence care delivery.

Acknowledgments

The authors would also like to acknowledge the support of Tres Brown from the Duke Health Technology System's Maestro Care Electronic Health Records (EHR) Application Team. This work was done while authors RS and TT were Clinical Informatics Fellows at Duke University.

This work was partially funded by a grant from the Duke Institute for Health Innovation.

Data Availability

The authors have uploaded the groupers to GitHub and will share if emailed directly.

Authors' Contributions

All the authors of this manuscript participated in and contributed equally to the conceptualization, design, and evaluation of the program list (PL) grouper categorization. EHM used the previous work of Heidi Twedt, MD, at Stanford (172 concepts across 19 systems and conditions) to develop the final set by extending 8 systems' coverage (eg, "Pulmonary" to "Pulmonary and Allergy"), adding 2 system groupers, and extending them to a total of 1211 concepts with the Boolean logic. Authors RS and EMH primarily authored the manuscript with other authors contributing to editing. Epic Systems incorporated these groupers into their standard development for PL organization by system as default in November of 2022. EMH will continue to make yearly updates to the groupers.

Conflicts of Interest

None declared.

References

1. Weed LL. Medical records that guide and teach. N Engl J Med 1968 Mar 14;278(11):593-600. [doi: 10.1056/NEJM196803142781105] [Medline: 5637758]
2. Wright A, Sittig DF, McGowan J, Ash JS, Weed LL. Bringing science to medicine: an interview with Larry Weed, inventor of the problem-oriented medical record. J Am Med Inform Assoc 2014 Dec;21(6):964-968. [doi: 10.1136/amiajnl-2014-002776] [Medline: 24872343]
3. 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. [doi: 10.1016/j.ijmedinf.2021.104452] [Medline: 33864979]
4. Henricks WH. “Meaningful use" of electronic health records and its relevance to laboratories and pathologists”. J Pathol Inform 2011 Feb 11;2:7. [doi: 10.4103/2153-3539.76733] [Medline: 21383931]
5. Wright A, McCoy AB, Hickman TT, et al. Problem list completeness in electronic health records: a multi-site study and assessment of success factors. Int J Med Inform 2015 Oct;84(10):784-790. [doi: 10.1016/j.ijmedinf.2015.06.011] [Medline: 26228650]
6. International statistical classification of diseases and related health problems (ICD). World Health Organization. URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2023-02-23]
7. SNOMED international recognizes entity linking challenge winners. SNOMED. URL: https://www.snomed.org/ [accessed 2023-02-23]
8. Moriyama IM, Loy RM, Robb-smith AHT, Rosenberg HM, Hoyert DL. History of the Statistical Classification of Diseases and Causes of Death: National Center for Health Statistics; 2011.
9. International Classification of Diseases, Tenth Revision (ICD-10). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/icd/icd10.htm [accessed 2023-02-23]
10. 2.05: ICD-10-CM. MedicalBillingandCoding.org. URL: https://www.medicalbillingandcoding.org/icd-10-cm/ [accessed 2023-03-14]
11. Alharbi MA, Isouard G, Tolchard B. Historical development of the statistical classification of causes of death and diseases. Cogent Med 2021 Jan 1;8(1):1893422. [doi: 10.1080/2331205X.2021.1893422]
12. 2024 ICD-10-CM. Centers for Medicare and Medicaid Services. URL: https://www.cms.gov/medicare/coding-billing/icd-10-codes/2024-icd-10-cm [accessed 2024-04-20]
13. Overview of SNOMED CT. National Library of Medicine. URL: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html [accessed 2023-03-14]
14. Davidso D, Rawson M. SNOMED CT: why it matters to you. Wolters Kluwer. URL: https://www.wolterskluwer.com/en/expert-insights/snomed-ct-why-it-matters-to-you [accessed 2023-02-23]
15. 5-step briefing. SNOMED International. URL: https://www.snomed.org/five-step-briefing [accessed 2023-02-24]
16. Release summary. SNOMED CT Release Statistics 2024-04-01. URL: https://browser.ihtsdotools.org/qa/#/SNOMEDCT/release-summary [accessed 2024-04-20]
17. SNOMED CT to ICD-10-CM map. US National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html [accessed 2023-03-02]
18. March 2021 - SNOMED CT managed service - US edition (US). SNOMED Confluence. 2021. URL: https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=121867053 [accessed 2023-03-02]

19. Klappe ES, de Keizer NF, Cornet R. Factors influencing problem list use in electronic health records-application of the unified theory of acceptance and use of technology. Appl Clin Inform 2020 May;11(3):415-426. [doi: 10.1055/s-0040-1712466] [Medline: 32521555]
20. Kreuzthaler M, Pfeifer B, Vera Ramos JA, et al. EHR problem list clustering for improved topic-space navigation. BMC Med Inform Decis Mak 2019 Apr 4;19(Suppl 3):72. [doi: 10.1186/s12911-019-0789-9] [Medline: 30943968]
21. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012 Oct;22(3):276-282. [doi: 10.11613/BM.2012.031] [Medline: 23092060]
22. Willett DL, Kannan V, Chu L, 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. [doi: 10.1055/s-0038-1668090] [Medline: 30157499]
23. Chu L, Kannan V, Basit MA, et al. SNOMED CT concept hierarchies for computable clinical phenotypes from electronic health record data: comparison of Intensional versus extensional value sets. JMIR Med Inform 2019 Jan 16;7(1):e11487. [doi: 10.2196/11487] [Medline: 30664458]
24. The CORE Problem List Subset of SNOMED CT®. US National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html [accessed 2023-03-18]
25. Fung KW, Xu J. Synergism between the mapping projects from SNOMED CT to ICD-10 and ICD-10-CM. AMIA Annu Symp Proc 2012;2012:218-227. [Medline: 23304291]
26. McGlothlin S. SNOMED and ICD: aligning to standards. J2 Interactive. 2021. URL: https://www.j2interactive.com/blog/snomed-and-icd/ [accessed 2023-03-02]
27. Fung KW, Xu J, Ameye F, Gutierrez AR, D'Have A. Achieving logical equivalence between SNOMED CT and ICD-10-PCS surgical procedures. AMIA Annu Symp Proc 2018 Apr 16;2017:724-733. [Medline: 29854138]
28. ICD-11 2022 release. World Health Organization. 2022. URL: https://www.who.int/news/item/11-02-2022-icd-11-2022-release [accessed 2023-03-15]

Abbreviations

EHR: electronic health record
HITECH: Health Information Technology Economic and Clinical Health
***ICD*:** *International Classification of Diseases*
IHTSDO: International Health Terminology Standards Development Organization
IMO: Intelligent Medical Objects
NEC: Not Elsewhere Classifiable
NOS: Not Otherwise Specified
PL: problem list
SNOMED: Systematized Nomenclature of Medicine
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SOAP: Subjective, Objective, Assessment, and Plan

Edited by C Lovis; submitted 27.07.23; peer-reviewed by C Gaudet-Blavignac, L Chu, T Karen; revised version received 01.12.23; accepted 22.02.24; published 09.05.24.

Please cite as:
Senior R, Tsai T, Ratliff W, Nadler L, Balu S, Malcolm E, McPeek Hinz E
Evaluation of SNOMED CT Grouper Accuracy and Coverage in Organizing the Electronic Health Record Problem List by Clinical System: Observational Study
JMIR Med Inform 2024;12:e51274
URL: https://medinform.jmir.org/2024/1/e51274
doi:10.2196/51274

Original Paper

Dermoscopy Differential Diagnosis Explorer (D3X) Ontology to Aggregate and Link Dermoscopic Patterns to Differential Diagnoses: Development and Usability Study

Rebecca Z Lin[1], MD; Muhammad Tuan Amith[2,3,4], PhD; Cynthia X Wang[5], MPHS, MD; John Strickley[6], MD; Cui Tao[7], PhD

[1]Division of Dermatology, Washington University School of Medicine, St. Louis, MO, United States
[2]Department of Information Science, University of North Texas, Denton, TX, United States
[3]Department of Biostatistics and Data Science, The University of Texas Medical Branch, Galveston, TX, United States
[4]Department of Internal Medicine, The University of Texas Medical Branch, Galveston, TX, United States
[5]Department of Dermatology, Kaiser Permanente Redwood City Medical Center, Redwood City, CA, United States
[6]Division of Dermatology, University of Louisville, Louisville, KY, United States
[7]Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL, United States

Corresponding Author:
Cui Tao, PhD
Department of Artificial Intelligence and Informatics
Mayo Clinic
4500 San Pablo Road
Jacksonville, FL, 32224
United States
Phone: 1 9049530255
Email: Tao.Cui@mayo.edu

Abstract

Background: Dermoscopy is a growing field that uses microscopy to allow dermatologists and primary care physicians to identify skin lesions. For a given skin lesion, a wide variety of differential diagnoses exist, which may be challenging for inexperienced users to name and understand.

Objective: In this study, we describe the creation of the dermoscopy differential diagnosis explorer (D3X), an ontology linking dermoscopic patterns to differential diagnoses.

Methods: Existing ontologies that were incorporated into D3X include the elements of visuals ontology and dermoscopy elements of visuals ontology, which connect visual features to dermoscopic patterns. A list of differential diagnoses for each pattern was generated from the literature and in consultation with domain experts. Open-source images were incorporated from DermNet, Dermoscopedia, and open-access research papers.

Results: D3X was encoded in the OWL 2 web ontology language and includes 3041 logical axioms, 1519 classes, 103 object properties, and 20 data properties. We compared D3X with publicly available ontologies in the dermatology domain using a semiotic theory–driven metric to measure the innate qualities of D3X with others. The results indicate that D3X is adequately comparable with other ontologies of the dermatology domain.

Conclusions: The D3X ontology is a resource that can link and integrate dermoscopic differential diagnoses and supplementary information with existing ontology-based resources. Future directions include developing a web application based on D3X for dermoscopy education and clinical practice.

(JMIR Med Inform 2024;12:e49613) doi:10.2196/49613

KEYWORDS

medical informatics; biomedical ontology; ontology; ontologies; vocabulary; OWL; web ontology language; skin; semiotic; web app; web application; visual; visualization; dermoscopic; diagnosis; diagnoses; diagnostic; information storage; information retrieval; skin lesion; skin diseases; dermoscopy differential diagnosis explorer; dermatology; dermoscopy; differential diagnosis; information storage and retrieval

Introduction

Dermoscopy is a noninvasive, in vivo microscopic technique used to examine skin lesions by detecting morphological features that may not be seen by the naked eye [1-4]. Studies have demonstrated that dermoscopy improves the diagnosis of both pigmented skin lesions [3,5-7] and nonpigmented skin lesions [8], including neoplasms [9] and infectious and inflammatory skin diseases [4,10]. Notably, the diagnostic accuracy of dermoscopy is dependent on the examiner's experience, as dermoscopy by untrained or less experienced examiners was found to be no better than clinical inspection without dermoscopy [6]. Learning dermoscopy is not just relevant to dermatologists, but also for physicians in other medical specialties. Patients with new or changing skin lesions often first consult their primary care physician (PCP) rather than a dermatologist. Dermoscopy has shown to be an effective tool for the assessment and triage of pigmented skin lesions in primary care, with improved diagnostic accuracy and referral accuracy to dermatologists [11-13]. However, dermoscopy training for PCPs is currently highly variable, with many PCPs citing a lack of training as a key barrier to the use of dermoscopy [14-16]. Furthermore, short dermoscopy training programs [14] may be insufficient to establish long-term competency in dermoscopy, with poor continuing use of dermoscopy and the need for refresher sessions [17]. The need for dermoscopy training among plastic surgeons has recently been documented as well [18]. Thus, the development of machine-based tools for dermoscopy may enhance clinical practice for dermatology providers and other medical professionals.

The use of standard terminologies organized through taxonomies has a long history with the life sciences, starting with Carl Linnaeus' taxonomy [19]: a classification system to name and group species according to their shared characteristics. Centuries later and with advances in computing infrastructure, these types of classification systems have continued to be of interest to the science community. An ontology is "a representational artifact comprising a taxonomy as proper part, whose representational units are intended to designate some combination of universals, defined classes, and certain relations between them" [20]. Essentially, an ontology is a graphical representation of linked concepts to formalize a schema (Tbox) for data (Abox). The formalization leverages semantic links (Rbox) between the concepts to give data more meaning and to aggregate related data of any heterogeneous format. This ensures the normalization of heterogeneous data. Furthermore, with semantics, ontologies could support machine reasoning to generate references via deductive reasoning. As related to the medical field, ontologies can extend the computability of standard controlled terminologies to provide descriptive and composite representations of medical information (such as features related to various diagnoses). Ontologies represent the data in a machine-readable format to give computing tools more context, making them highly valuable for artificial intelligence.

Within the dermatology domain, some existing ontologies aim to describe cutaneous disorders. For example, the dermatology lexicon (DERMLEX) was created with the American Academy of Dermatology with a nosology, anatomical distributions, classical signs, and therapeutic procedures; however, maintenance was discontinued in 2009 [21,22]. More recently, the human dermatological disease ontology (DERMO) was developed to classify cutaneous diseases by etiology, anatomical location or cell type, and phenotype consistent with current clinical practice [23,24]. Some other dermatology-specific ontologies exist, including the skin physiology ontology (SPO; last updated in 2008) [25], but notably, none of these ontologies connect cutaneous disorders to metaphoric terms like "strawberry pattern" which may be difficult for a machine to understand. Similarly, none of the aforementioned ontologies specifically address dermoscopy, which is a specialized technique that may have special considerations when used in diagnosis. For instance, the colors of certain lesions are best seen under polarized light [26]. As such, there is a need to develop an ontology that adequately addresses the field of dermoscopy, with the capability of processing both descriptive and metaphoric terminology.

In our previous work, we developed the elements of visuals ontology (EVO) to decompose the fundamental features of visualizations, such as shapes, colors, and textures. The dermoscopy elements of visuals ontology (DEVO) then applied the visual features described in EVO to dermoscopic terminology [27]. For instance, DEVO characterizes dermoscopic metaphoric terms such as "shiny white streaks" and "leaflike areas" by shapes, colors, and textures, along with other features involved. Discussion with domain experts revealed that while DEVO is capable of responding to queries to find visual features associated with metaphoric terms and vice versa, linking the dermoscopic terms to differential diagnoses would significantly enhance its clinical utility. A list of differential diagnoses indicates many possible diagnoses that share similar features to the patient's symptoms and signs. These differential diagnoses can then be narrowed down to aid the clinician in identifying the final diagnosis. As dermatology is a technical field, the landscape of differential diagnoses is wide and difficult to parse [28]. In this study, we describe the extension of EVO and DEVO to create the dermoscopy differential diagnosis explorer (D3X), an ontology linking metaphoric terms to differential diagnoses. We further propose a use case integrating D3X into a web application in dermoscopy education and clinical practice.

Methods

Ethical Considerations

This article adheres to the Committee on Publication Ethics guidelines. This research did not involve human subjects.

Integration of Existing Ontologies

Overview

A common practice in the development of ontologies [29] is to reuse existing ontologies' components to ensure semantic interoperability. We used the following ontologies to build the D3X ontology.

About EVO

EVO is a foundational ontology model that describes the basic constituents of visualizations: shapes, colors, strokes (lines), size, perceived texture, etc. It also represents the dimensional extended 9-intersection model, a mathematical model for spatial relationships between elements [30]. Further, EVO imports and reuses controlled terminologies and standards from the W3C scalable vector graphics, Wikidata, phenotype and trait ontology, and the simple knowledge organization system to supplement our core representational model of visualizations. EVO is hosted on GitHub for public release and is coded in the OWL 2 web ontology language.

About DEVO

DEVO is an extension of EVO that reuses the foundational understanding of visualizations for the dermoscopy domain. DEVO incorporates some of the controlled terminologies—"metaphoric" and "descriptive"—that are used in practice by dermatologists, with a focus on the metaphoric terminologies. With DEVO, we developed a core model that encodes and describes the "visual language" of the dermoscopic terms' definitions. Further, one important outcome of this work was a computable representational model of an agreed understanding of visual elements of dermoscopic patterns, which we used as a framework to generate differential diagnoses. Similarly, DEVO was coded in OWL 2 and is hosted on GitHub for public consumption.

Miscellaneous Ontologies and Vocabularies

We also aligned D3X with commonly used top-level ontologies. The information artifact ontology (IAO) [31] is part of the open biological and biomedical ontology (OBO) foundry. IAO represents a general abstraction of informational objects (like documents and components within those documents—eg, figures, images). Like many OBO foundry ontologies, IAO uses the basic formal ontology and relation ontology as part of its architecture model. We minimally reused some of the term entities and properties like IAO:image and "denoted by." We also reused the software ontology (SWO) [32] for its licensing entity terms—SWO:license and "has license"—to describe the licensing information for any imaging resource of skin lesions. Lastly, we used Schema.org's [33] schema::image to link image resources.

Development of D3X

To generate a list of differential diagnoses, we started with the metaphoric terms defined in DEVO from the third consensus conference conducted by the International Society of Dermoscopy [34]. We then searched the literature [34-36] for corresponding differential diagnoses for each term and consulted 2 domain experts to independently edit the list of diagnoses for accuracy. These differential diagnoses were later encoded using Protégé [37] in our ontology. Following this, we reviewed open-source resources (DermNet, Dermoscopedia, and open-access research papers) for a collection of hosted images that matched individual differential diagnoses. We tracked the provenance information and associated data (caption, description, etc) in a spreadsheet as a central organized resource that mapped the images for each diagnosis to the concept diagnosis used in D3X. To streamline the data transfer process, we developed a management code to transfer data from the spreadsheet to the ontology. The source code is available on our GitHub repository, using the OWL API to facilitate efficient custom import. This approach allowed centralized data collection and also enabled an ad-hoc import and data creation pipeline.

Semiotic Evaluation

Semiotic theory is the study of signs and symbols, and considering ontologies are symbolic representations of a specific domain, we used a metric suite grounded in that theoretical framework [38]. Semiotic theory is composed of 3 basic qualities: *syntactic*, *semantic*, and *pragmatic*. Essentially, in the context of ontologies, the metric suite components refer to aspects of the ontology artifact—*syntactic* concerning encoding adherence and standards; *semantic* concerning the effective use of human-friendly labels for entities and concepts; and *pragmatic* concerning function. Each of these qualities is quantified based on a computation of representative quantifiable features of an ontology file (eg, the number of classes, the average number of word senses for labels, etc). This is described in detail in previously published works [38]. This suite helps to measure some of the intrinsic qualities of our ontology concerning other ontologies in the same domain. We used publicly available ontologies from the skin and dermatology domain—DERMLEX, DERMO, and SPO—that are found in the National Center for Biomedical Ontology (NCBO) BioPortal. We used a command line version of our tool OntoKeeper [39] to quickly generate scores from the metric suite and then calculated *z* scores to determine how D3X fares in terms of intrinsic quality with other ontologies of the dermatology domain.

Results

Development of D3X

The D3X ontology was encoded in the OWL 2 web ontology language. In terms of the size of the ontology, there are 3041 logical axioms, 1519 classes, 103 object properties, and 20 data properties. Imported image data are encoded as 387 instances. Figure 1 displays a sample series of screenshots showing Kaposi sarcoma, as an example entity, linked to DEVO's rainbow pattern, standard medical terminologies (eg, Systematized Nomenclature of Medicine–Clinical Terms [SNOMED CT], National Cancer Institute [NCI] Thesaurus), and the open-sourced image example. For ongoing data management, we host the spreadsheet with image data (n=364 images) and the OWL API software code to allow for an automated process of adding new image data. The software will pull the data from the spreadsheet and will add and export a version of our ontology that has the image instance data. Both the spreadsheet and the software are available on our GitHub repository [32]. As more dermoscopy images become available for the public domain, we will include them in our spreadsheet and generate an encoded export with the new instance data. D3X uses our pre-existing work of DEVO and also leverages terminology from the IAO, SWO, and Schema.org. Figure 2 shows a global

overview of the D3X ontology and the various linked terminologies that compose the entire model.

Figure 1. Sample screenshot of the D3X ontology through Protégé showing related metadata and information about Kaposi sarcoma. D3X: dermoscopy differential diagnosis explorer; DDX: differential diagnosis.

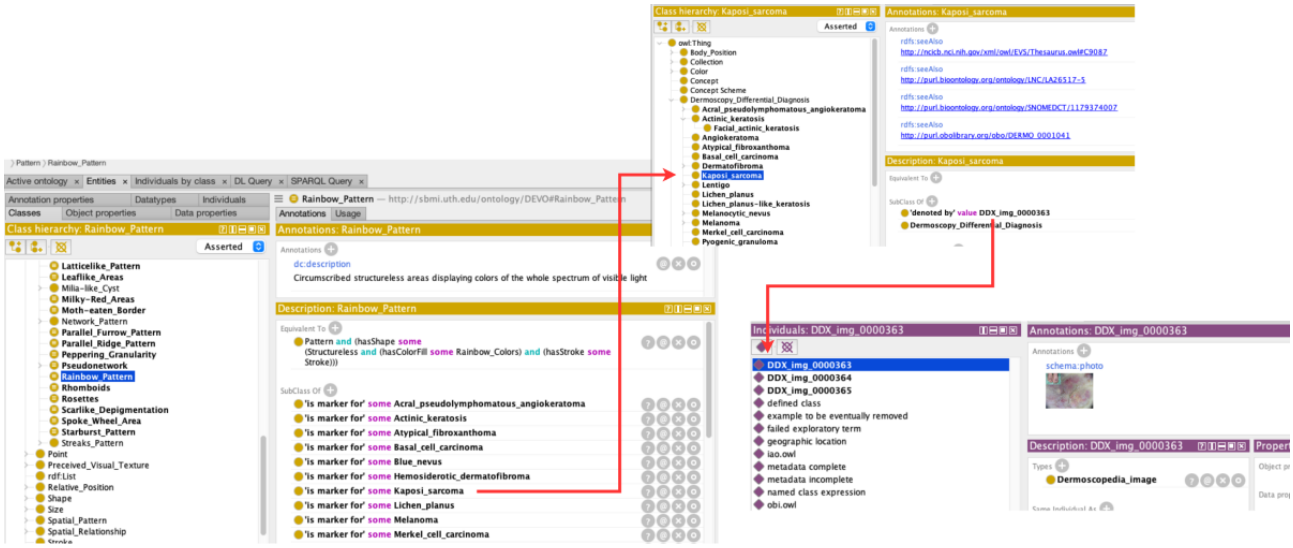
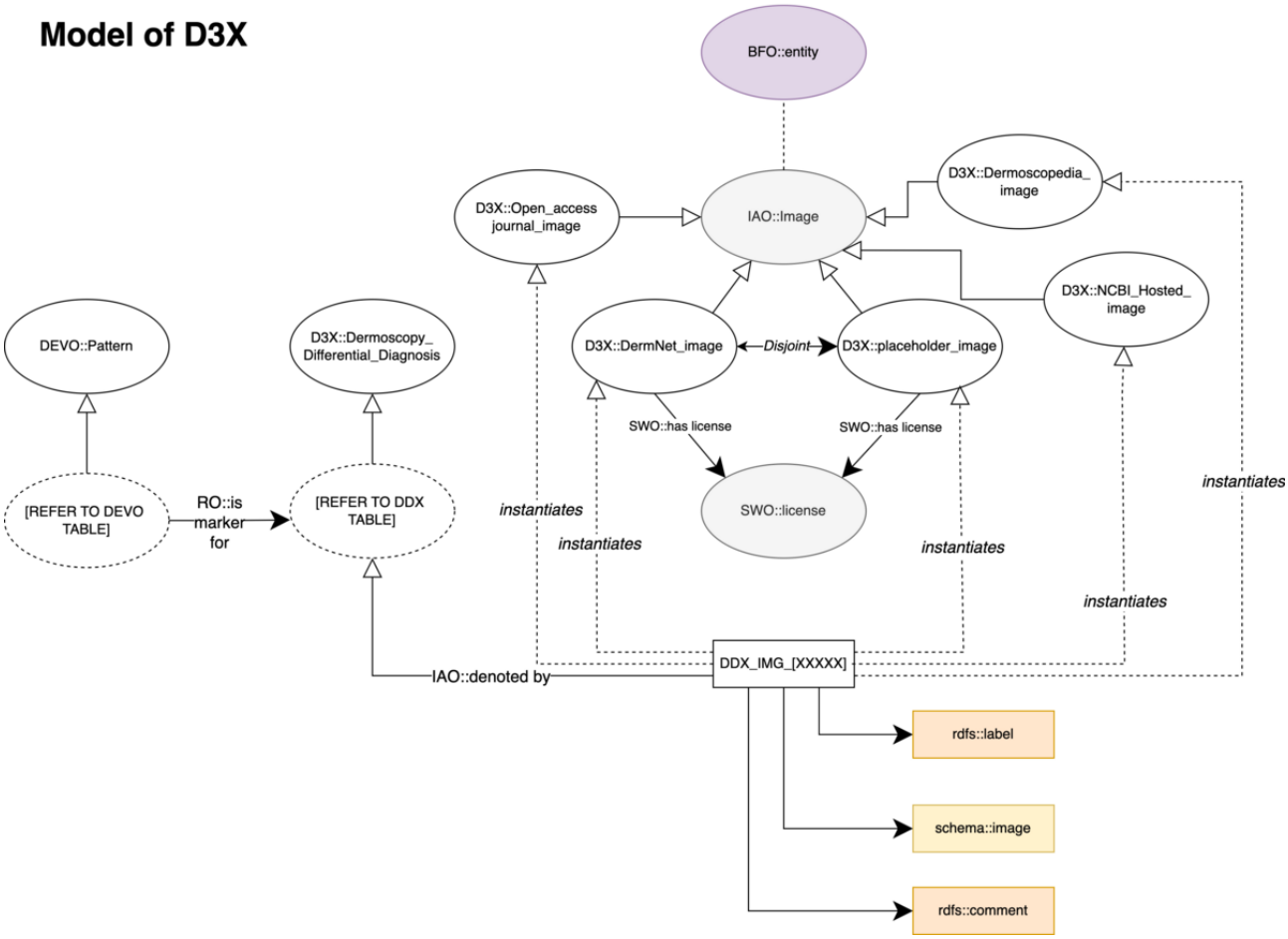


Figure 2. Global overview of the D3X ontology and linked ontologies, including the DEVO, BFO, IAO, and SWO. BFO: basic formal ontology; D3X: dermoscopy differential diagnosis explorer; DEVO: dermoscopy elements of visuals ontology; IAO: information artifact ontology; rdfs: resource description framework schema; RO: relation ontology; SWO: software ontology.



Each dermoscopy sample image is represented as a single instance data value with a unique ID (DDX_IMG_[DIGITS]). As an instance data value representing the digital image, it links to the exact file on the web using schema:image from Schema.org. Caption information is used as an annotation for RDF:comment (RDF: resource description framework) and rdf:label (rdf: resource description framework). The instance data value is an instantiation of a specific image class from the

following sources: IAO—open access journal images, DermNet images, National Center for Biotechnology Information–hosted images, and Dermoscopedia images. For some of them, the "has licenses" predicate links to a license, signifying that any instance of this image class has some license agreement (eg, Creative Commons). The licensing terminology is derived from the SWO and is hosted on our GitHub repository as an external import.

With D3X, we declared a new dermoscopy differential diagnosis class. This class provides a list of associated diagnoses for skin lesions. Each of the dermoscopy differential diagnosis classes is linked to a pattern from DEVO. The pattern in DEVO ontologically describes each dermoscopic pattern (metaphoric term) using visual elements, such as lines, shapes, colors, and spatial relationships. Table S1 in Multimedia Appendix 1 provides a comprehensive list of the metaphoric patterns listed in DEVO and their corresponding differential diagnoses in D3X. Each pattern in DEVO is linked to its differential diagnoses using OBO's "is marker for" (eg, angular lines > is a marker for > Lentigo_maligna). Additionally, the instance images described above are linked to the differential diagnoses using "denoted by," such that each diagnosis is provided with at least one visual example. Lastly, for each of the dermoscopy differential diagnosis classes, there are associated annotations that link the class to the other standardized ontologies like the Medical Dictionary for Regulatory Activities (MedDRA), SNOMED CT, NCI Thesaurus, and LOINC (logical observation identifier names and codes). MedRA covered 63% (n=25) of the classes, while SNOMED CT and NCI Thesaurus covered 53% (n=21) and 55% (n=22) of the classes, respectively. The remaining, like DERMO and LOINC, covered 15% (n=6) and 3% (n=1) of the classes.

Semiotic Evaluation

Semiotic theory is composed of 3 basic qualities: *syntactic*, *semantic*, and *pragmatic*. Table 1 displays the *z* scores for each of the qualities and subqualities of D3X compared to other publicly available ontologies in the dermatology domain. Examining the *syntactic* quality of D3X (z=0.17), while it lacks diverse syntactic *richness* (z=–0.74) in comparison with its other domain counterparts, D3X does adhere to syntactic *lawfulness* (z=0.49). D3X compares satisfactorily with other ontologies in the *semantic* quality (z=0.77). Although the semantic *clarity* subquality was below average than its peers (z=–0.91; the ambiguity of labels), D3X does better with semantic *consistency* (z=0.56; the number of essentially unique labels) and semantic *interpretability* (z=0.65; whether the label has meaning). The *pragmatic* quality is composed of 1 score: *comprehensiveness*, a measure of the coverage of the domain scope of the ontology based on the number of entities encoded, which was nearly below average for D3X (z=–0.66). Lastly, the overall score of D3X (z=0.58) points to a somewhat better overall quality score than DERMLEX and SPO (z=–1.41 and 0.00, respectively). Although DERMO had a slightly higher overall quality than D3X (z=0.83), its score is still within 1 SD of the D3X ontology score, so the quality of D3X appears at least comparable to that of the other ontologies within its own domain.

Table 1. Semiotic comparison of D3X[a] to other dermatology ontologies: the DERMLEX[b], DERMO[c], and SPO[d] using *z* scores.

Quality and subquality	Mean (SD)	D3X-z	DERMLEX-z	DERMO-z	SPO-z
Syntactic	0.57 (0.11)	0.17[e]	–1.33	0.08	1.09[e]
Richness	0.26 (0.11)	–0.74	0.62[e]	–0.96	1.08[e]
Lawfulness	0.87 (0.25)	0.49	–1.50	0.51[e]	0.51 [e]
Semantic	0.85 (0.13)	0.77[e]	–1.38	0.70[e]	–0.08
Clarity	0.99 (0.01)	–0.91	0.91[e]	0.82[e]	–0.82
Consistency	0.73 (0.49)	0.56[e]	–1.50	0.43	0.51[e]
Interpretability	0.87 (0.21)	0.65[e]	0.65[e]	0.16	–1.46
Pragmatic	0.11 (0.15)	–0.66	1.44[e]	–0.09[e]	–0.69
Comprehensiveness	0.11 (0.15)	–0.66	1.44[e]	–0.09[e]	–0.69
Overall score	0.52 (0.03)	0.58 [e]	–1.41	0.83[e]	0.00

[a]D3X: dermoscopy differential diagnosis explorer.

[b]DERMLEX: dermatology lexicon.

[c]DERMO: human dermatological disease ontology.

[d]SPO: skin physiology ontology.

[e]These values indicate the 2 highest values for each quality and subquality.

Discussion

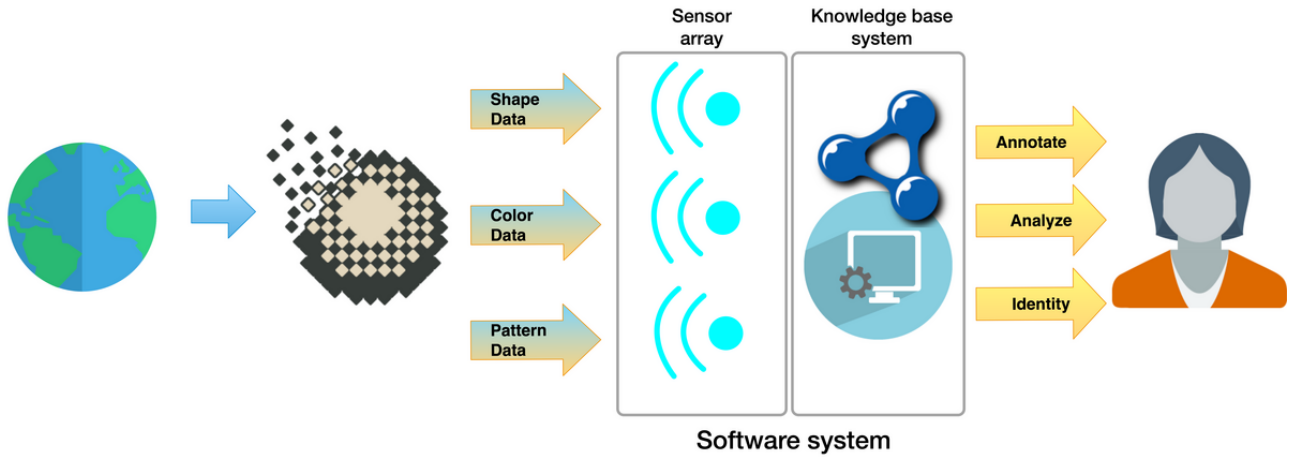
Principal Results and Limitations

D3X is an ontology that connects dermoscopic patterns (metaphoric terms) with differential diagnoses. It is an extension of the DEVO to describe patterns based on their visual elements, which is in turn an extension of the EVO. D3X also leverages terminology from IAO, SWO, Relation Ontology, and Schema.org, and its differential diagnoses are linked to MedDRA, SNOMED CT, and NCI Thesaurus. Using the semiotic theory framework proposed by Burton-Jones et al [38], we measured D3X in comparison with similar publicly available ontologies to assess its intrinsic quality. Our assessment indicates that while comparably better to the other ontologies of the same domain in its overall score, D3X does lack diverse syntactic *richness* and could improve its semantic *clarity* (despite a better overall semantic quality score than its ilk) and pragmatic *comprehensiveness*. Leveraging additional OWL 2 syntactic features could improve the syntactic richness. However, since the purpose of our ontology is to retrieve and aggregate information and metadata about dermoscopic features, some of the more sophisticated OWL 2 features like symmetry, inverse, etc, may not be necessary for our use case. As for the pragmatic score, it might improve over time as we collect more instance data of images to link to our ontology. Further, our assessment was limited to 3 ontologies as there are no other publicly available ontologies that deal solely with a dermatology subject. Additionally, OntoKeeper uses a subset of scores as the social quality (composed of *authority* and *history*), and the pragmatic subscores of *accuracy* and *relevancy* are difficult to compute, so they are not listed in our semiotic analysis [39]. Despite this, the quality scores are sufficient for an application ontology, since the role of this artifact is to aggregate and consolidate skin diagnostic information—an area where it is likely to shine.

The aforementioned evaluation included DERMLEX, DERMO, and the SPO. DERMLEX was originally created by the American Academy of Dermatology to describe dermatological diagnosis and related domain vocabularies, aligned to *International Classification of Diseases, Ninth Revision* (*ICD-9*). However, the upkeep ended in 2009 [22]. DERMO is another ontology that also aims to describe dermatological diseases, but unlike DERMLEX, it is aligned to *International Classification of Diseases, Tenth Revision* (*ICD-10*). The latest version was last released in 2015, according to the NCBO BioPortal record [23]. Not much is known about SPO, other than a presence on NCBO BioPortal and the latest release dating back to 2008 [25]. Compared to these existing works, D3X yields richer semantics and applicability by the OWL2 encoding in EVO and DEVO that describes lesions using primitive visualization elements. Another advantage of this work is the use of semantic web properties of our work, namely the linking of heterogeneous resources (external entities, images, metadata, etc). This allows D3X to be an application-driven artifact that can be integrated into software tools, and other analytical and educational tools. According to researchers, terminologies enriched with semantics will yield opportunities to develop innovative tools and applications [40]. We further discuss our vision in the subsequent sections (see Proposed Web Application and Use Case). Overall, we presume this work provides a richer ontological artifact compared to similar ontologies of the same domain.

Aside from our aforementioned application use case, this work can advance machine learning models for dermoscopy diagnosis support. There has been some preliminary evidence that machine learning models can be supported or improved by ontologies [41-43]. Potentially, the combined stack of EVO, DEVO, and D3X could augment tools that analyze real-world entities (eg, lesions). In Figure 3, we illustrate a hypothetical example where a software application segments signals from an entity using machine learning in a sensor array to detect shape, color, and pattern data. The structured information from the sensor array could then be linked to an ontological knowledge base system that expresses meaning and context.

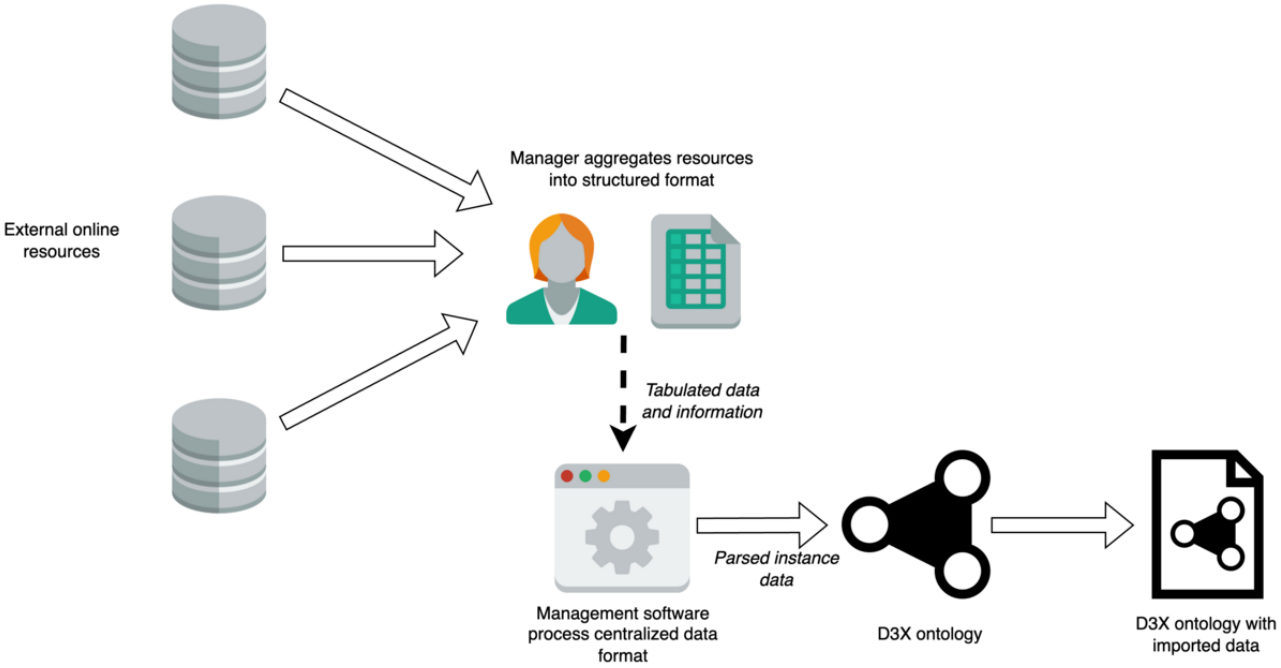
Figure 3. Diagram of a software system using segmented machine learning with a sensor array linked to an ontological knowledge base system. Pixel art icon by DesignContest is licensed under CC BY 4.0. Earth icon by Treetog ArtWork. Globe icon and user female icon by paomedia.



Data Upkeep and Management Plan

Noted earlier, we produced a basic management system to allow for continued integration of data and information from external sources to be added to D3X. Continued data management is an issue with some ontology and controlled terminology resources. By having this basic management system, we can ensure that D3X will be up to date with little resources and time needed to integrate new diagnostic information and metadata. Figure 4 shows the basic management pipeline, with the tools needed, hosted on our GitHub repository under the ddx_data_management folder. In the aforementioned figure, any new or updated digital resources (images, web page text, knowledge graph, and ontology resources) will be added to a centralized spreadsheet for the human-friendly organization of data for diagnosis information. The management software will import the spreadsheet and parse the data for the D3X ontology. The final output of the software is the D3X ontology with the updated linked information. Future plans could include using shapes constraint language (SHACL) to ensure the quality of the data is validated, and further development of data management software to facilitate ease of use.

Figure 4. Outline of the D3X ontology data upkeep and management plan. D3X: dermoscopy differential diagnosis explorer. OWL Lite icon and OWL Lite document icon by Picol Team are licensed under CC BY 4.0. File excel icon, database icon, window system icon, user female alt icon by paomedia.

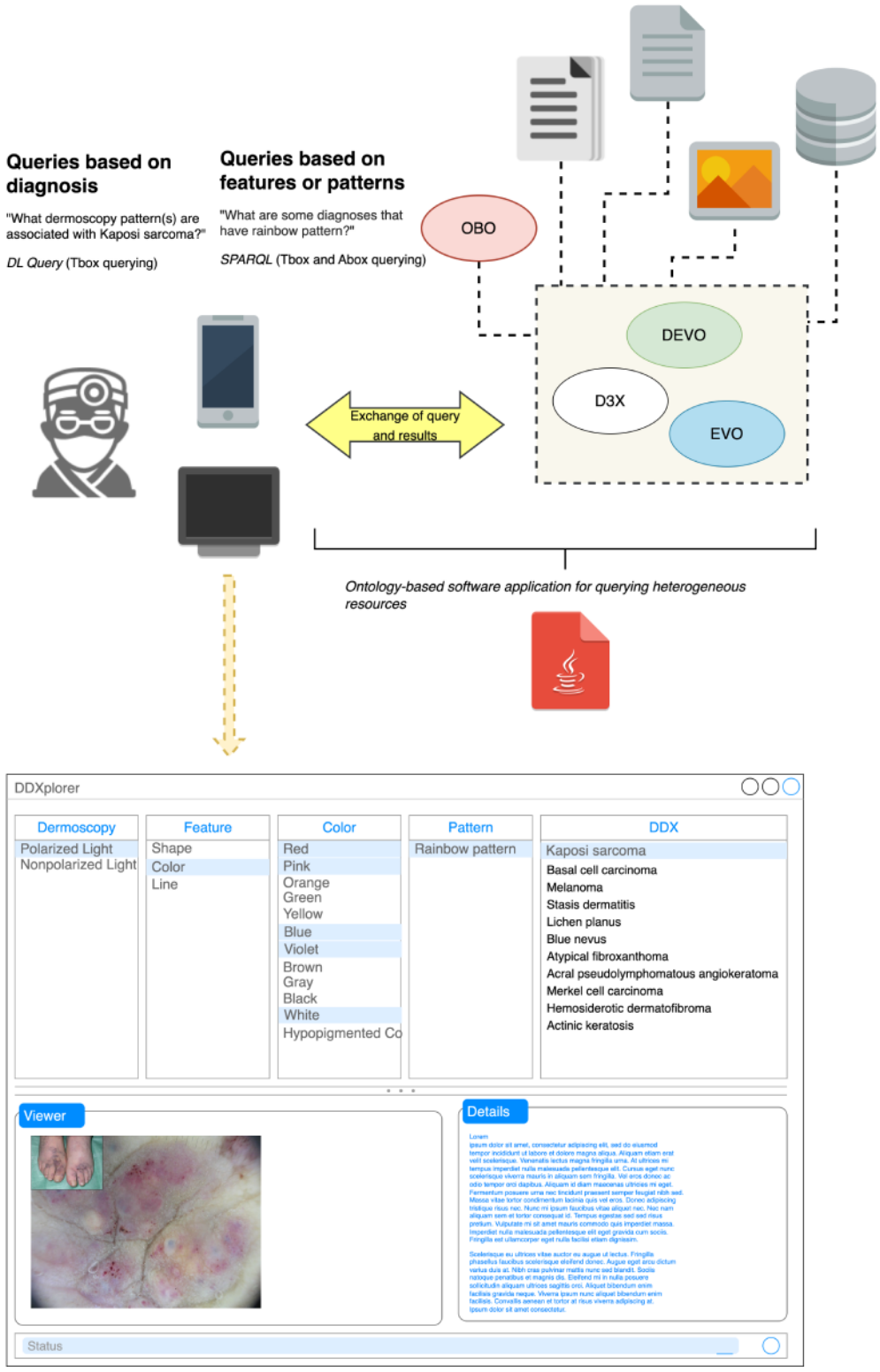


Proposed Web Application and Use Case

For use in clinical practice, an ontology-based software application could be designed using D3X to guide the identification of differential diagnoses. After the user performs dermoscopy on a skin lesion, they can open the web application and select various features about the lesion, which will then suggest differential diagnoses. A mock-up for the use case of Kaposi sarcoma is shown in Figure 5. The user can select options for the first 3 boxes ("dermoscopy," "feature," and the chosen feature [eg, "color"], with multi-select functionality available for the latter). The web application would then generate the relevant patterns and a list of differential diagnoses. Thus, if the user indicates that under polarized light, red, pink, blue, violet, and white colors were visualized, this corresponds to a "rainbow pattern," which is associated with Kaposi sarcoma, among other differential diagnoses. Clicking on each differential diagnosis will then display any relevant dermoscopic images in the "viewer," as well as a description under "details" with a link to learn more about the condition. By reviewing images and descriptions of these differential diagnoses, this would ideally help the user narrow down the list and identify the most likely diagnosis. This web application was independently reviewed by 2 domain experts who agreed that the format was understandable to the user; they also stated that the information provided would be useful in clinical practice as a quick search for differential diagnoses. Of note, the aforementioned example illustrates a query based on features or patterns, but the web application would also be capable of querying based on diagnoses (eg, "what dermoscopy patterns are associated with Kaposi sarcoma?"). This would be more useful in an educational setting for those who want to gain an understanding of dermoscopic patterns and the features comprising each pattern. Furthermore, we proposed the development of a web application harnessing D3X capable of carrying out the following queries: (1) given dermoscopic features or patterns, output a list of differential diagnoses; and (2) given a differential diagnosis, output associated features and patterns. Along with a description of each differential diagnosis, the application would also display images from DermNet, Dermoscopedia, National Center for Biotechnology Information Hosted, or open-access journals for ease of understanding the relationship between each differential diagnosis and its visual elements. There is a growing body of literature on machine learning models for automated diagnosis of dermoscopic images, such as convolutional neural networks (CNNs) [9]. Both ontologies and CNNs fall under the artificial intelligence umbrella, but ontologies relate to knowledge representation, while CNN is statistical machine learning. These

are fundamentally different approaches to power artificial intelligence that are difficult to compare directly. While automated diagnosis via CNNs is a very promising area of study, research has largely focused on the diagnosis of melanoma [44-46], with few studies including pigmented nonmelanocytic lesions [47,48] and largely ignoring nonpigmented lesions. D3X labels dermoscopic patterns of pigmented and nonpigmented lesions, so it may apply to a broader range of patient visits. Additionally, the likelihood of provider acceptance of automated diagnosis systems is unclear. With our proposed web application, providers would be able to input search criteria themselves and see a list of differential diagnoses, rather than a binary output for 1 diagnosis (eg, melanoma) suggested by the machine, which may not be as likely to be accepted by physicians.

Figure 5. Mock-up of a web application harnessing the D3X ontology to perform queries for differential diagnoses associated with dermoscopic features and patterns, with Kaposi sarcoma as a use case. Doctor Icon by MedicalWP is licensed under CC BY 4.0. Computer icon and text x java icon by Papirus Dev Team are licensed under GNU GPL (version 3.0). Database icon, file text icon, file picture icon, device mobile phone icon by paomedia. Abox: assertion component of a knowledge base; D3X: dermoscopy differential diagnosis explorer; DEVO: dermoscopy elements of visuals ontology; DL: description logic; EVO: elements of visuals ontology; OBO: open biological and biomedical ontology; SPARQL: SPARQL protocol and RDF query language (recursive acronym); Tbox: terminology component of a knowledge base.



In discussion with domain experts, we chose not to integrate diagnostic rules into D3X, as experienced dermoscopy users could assess the list of differential diagnoses fairly quickly and decide on the most likely diagnosis using their clinical expertise. If the user is relatively inexperienced, they may gain more understanding by reading the description of each differential diagnosis, viewing the images, and accessing additional information by clicking “learn more” in the web application. Nevertheless, in the future, it may be useful to integrate diagnostic rules into D3X for more sophisticated suggestions

of differential diagnoses (eg, ranking the most to least likely differential diagnoses in a prioritized list). Another limitation is that there are dermoscopic terms and differential diagnoses not included in D3X, given that we built D3X from the terms mentioned in the International Society of Dermoscopy's third consensus conference [34]. Similarly, while we aimed to include images from a variety of external sources, we acknowledge that they may not be fully representative of all patient skin tones. Our work is only a starting point, as we plan to continue updating D3X and anticipate that its library will become more comprehensive with time.

To conclude, the web application based on D3X has great potential for use in several areas. First, it could be included alongside formal dermoscopy training as a supplementary educational tool for dermatology trainees and providers in other specialties (PCPs, plastic surgeons). Given that providers may require ongoing dermoscopy refresher sessions to feel fully comfortable even after completing an initial training program [17], this web application could be a helpful reference to deepen understanding of dermoscopic patterns associated with different skin conditions. Furthermore, in a clinical setting, providers could quickly query the web application for a list of differential diagnoses after dermoscopic examination of a lesion, which would aid in the identification of their patient's diagnosis. This may be useful for inexperienced and experienced dermoscopy users alike, as it is intended to augment, not replace, the provider's clinical reasoning. The next steps include using our proposed interface to build a functional web application. Creating the web application could reveal additional flaws in the design that require clarification, and we would continue to improve aspects of D3X and the web application in an iterative process. After a beta version of the web application is finalized, we would aim to conduct user testing to evaluate the user experience as well as clinical or educational utility among physicians.

Conclusions

We introduce and discuss the design and development of the D3X ontology as a resource that can link and integrate dermoscopic differential diagnoses and supplementary information with existing ontology-based resources (MedDRA, SNOMED CT, and NCI). We repurposed a previous work of the EVO and DEVO to construct and support D3X, along with other supplementary standardized ontologies, like IAO and SWO. Using the semiotic theoretical framework to compare D3X with other dermatology-related ontologies, its overall quality score was similar to existing ontologies' scores. One of the outcomes of this work is providing a means to aggregate and link dermoscopic patterns to differential diagnoses, thereby enhancing understanding of dermoscopy for educational and clinical use. This outcome has fueled our next objective in developing a web-based application that can query D3X and fetch the linked information for the user. Currently, D3X and its resources are available on GitHub for public release and use.

Acknowledgments

This work was partially supported by the National Institutes of Health (R01AI130460 and U24CA194215) and the Cancer Prevention Research Institute of Texas (RP220244).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Patterns in the dermoscopy elements of visuals ontology and their associated differential diagnoses in the dermoscopy differential diagnosis explorer.
[DOCX File , 19 KB - medinform_v12i1e49613_app1.docx]

References

1. Argenziano G, Soyer HP. Dermoscopy of pigmented skin lesions—a valuable tool for early diagnosis of melanoma. Lancet Oncol 2001;2(7):443-449. [doi: 10.1016/s1470-2045(00)00422-8] [Medline: 11905739]
2. Vázquez-López F, Manjón-Haces JA, Maldonado-Seral C, Raya-Aguado C, Pérez-Oliva N, Marghoob AA. Dermoscopic features of plaque psoriasis and lichen planus: new observations. Dermatology 2003;207(2):151-156. [doi: 10.1159/000071785] [Medline: 12920364]
3. Rosendahl C, Tschandl P, Cameron A, Kittler H. Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. J Am Acad Dermatol 2011;64(6):1068-1073. [doi: 10.1016/j.jaad.2010.03.039] [Medline: 21440329]
4. Lallas A, Argenziano G, Apalla Z, Gourhant JY, Zaballos P, Di Lernia V, et al. Dermoscopic patterns of common facial inflammatory skin diseases. J Eur Acad Dermatol Venereol 2014;28(5):609-614. [doi: 10.1111/jdv.12146] [Medline: 23489377]
5. Bafounta ML, Beauchet A, Aegerter P, Saiag P. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. Arch Dermatol 2001;137(10):1343-1350. [doi: 10.1001/archderm.137.10.1343] [Medline: 11594860]
6. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. Lancet Oncol 2002;3(3):159-165. [doi: 10.1016/s1470-2045(02)00679-4] [Medline: 11902502]

7. Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Br J Dermatol 2008;159(3):669-676. [doi: 10.1111/j.1365-2133.2008.08713.x] [Medline: 18616769]
8. Sinz C, Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, et al. Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. J Am Acad Dermatol 2017;77(6):1100-1109. [doi: 10.1016/j.jaad.2017.07.022] [Medline: 28941871]
9. Weber P, Tschandl P, Sinz C, Kittler H. Dermatoscopy of neoplastic skin lesions: recent advances, updates, and revisions. Curr Treat Options Oncol 2018;19(11):56 [FREE Full text] [doi: 10.1007/s11864-018-0573-6] [Medline: 30238167]
10. Haliasos EC, Kerner M, Jaimes-Lopez N, Rudnicka L, Zalaudek I, Malvehy J, et al. Dermoscopy for the pediatric dermatologist part I: dermoscopy of pediatric infectious and inflammatory skin lesions and hair disorders. Pediatr Dermatol 2013;30(2):163-171. [doi: 10.1111/pde.12097] [Medline: 23405886]
11. Westerhoff K, McCarthy WH, Menzies SW. Increase in the sensitivity for melanoma diagnosis by primary care physicians using skin surface microscopy. Br J Dermatol 2000;143(5):1016-1020. [doi: 10.1046/j.1365-2133.2000.03836.x] [Medline: 11069512]
12. Argenziano G, Puig S, Zalaudek I, Sera F, Corona R, Alsina M, et al. Dermoscopy improves accuracy of primary care physicians to triage lesions suggestive of skin cancer. J Clin Oncol 2006;24(12):1877-1882. [doi: 10.1200/JCO.2005.05.0864] [Medline: 16622262]
13. Herschorn A. Dermoscopy for melanoma detection in family practice. Can Fam Physician 2012;58(7):740-745 [FREE Full text] [Medline: 22859635]
14. Chappuis P, Duru G, Marchal O, Girier P, Dalle S, Thomas L. Dermoscopy, a useful tool for general practitioners in melanoma screening: a nationwide survey. Br J Dermatol 2016;175(4):744-750. [doi: 10.1111/bjd.14495] [Medline: 26914613]
15. Morris JB, Alfonso SV, Hernandez N, Fernández MI. Examining the factors associated with past and present dermoscopy use among family physicians. Dermatol Pract Concept 2017;7(4):63-70 [FREE Full text] [doi: 10.5826/dpc.0704a13] [Medline: 29214111]
16. Fee JA, McGrady FP, Rosendahl C, Hart ND. Training primary care physicians in dermoscopy for skin cancer detection: a scoping review. J Cancer Educ 2020;35(4):643-650 [FREE Full text] [doi: 10.1007/s13187-019-01647-7] [Medline: 31792723]
17. Robinson JK, MacLean M, Reavy R, Turrisi R, Mallett K, Martin GJ. Dermoscopy of concerning pigmented lesions and primary care providers' referrals at intervals after randomized trial of mastery learning. J Gen Intern Med 2018;33(6):799-800 [FREE Full text] [doi: 10.1007/s11606-018-4419-5] [Medline: 29637481]
18. Brennan MC, Kabuli MN, Dargan MD, Pinder MR. A short correspondence piece to the editor in chief: the need for increased training in the technique of dermoscopy amongst plastic surgeons and the under recognised value of dermoscopy in the assessment of non-pigmented cutaneous lesions. J Plast Reconstr Aesthet Surg 2022;75(1):496-498. [doi: 10.1016/j.bjps.2021.11.014] [Medline: 34852970]
19. Paterlini M. There shall be order. The legacy of Linnaeus in the age of molecular biology. EMBO Rep 2007;8(9):814-816 [FREE Full text] [doi: 10.1038/sj.embor.7401061] [Medline: 17767191]
20. Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a reference terminology for ontology research and development in the biomedical domain. 2006 Presented at: Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2006) ;222?65; November 8, 2006; Baltimore, MA p. 57.
21. Papier A, Chalmers RJG, Byrnes JA, Goldsmith LA, Dermatology Lexicon Project. Framework for improved communication: the Dermatology Lexicon Project. J Am Acad Dermatol 2004;50(4):630-634. [doi: 10.1016/s0190-9622(03)01571-8] [Medline: 15034516]
22. Dermatology lexicon. NCBO BioPortal. 2009. URL: https://bioportal.bioontology.org/ontologies/DERMLEX/?p=summary [accessed 2024-01-05]
23. Fisher HM, Hoehndorf R, Bazelato BS, Dadras SS, King LE, Gkoutos GV, et al. DermO; an ontology for the description of dermatologic disease. J Biomed Semantics 2016;7:38 [FREE Full text] [doi: 10.1186/s13326-016-0085-x] [Medline: 27296450]
24. Human dermatological disease ontology. NCBO BioPortal. URL: https://bioportal.bioontology.org/ontologies/DERMO [accessed 2024-01-05]
25. Skin physiology ontology. NCBO BioPortal. URL: https://bioportal.bioontology.org/ontologies/SPO [accessed 2024-01-05]
26. Benvenuto-Andrade C, Dusza SW, Agero ALC, Scope A, Rajadhyaksha M, Halpern AC, et al. Differences between polarized light dermoscopy and immersion contact dermoscopy for the evaluation of skin lesions. Arch Dermatol 2007;143(3):329-338. [doi: 10.1001/archderm.143.3.329] [Medline: 17372097]
27. Lin R, Amith M, Zhang X, Wang C, Light J, Strickley J, et al. Developing ontologies to standardize descriptions of visual and dermoscopic elements. 2021 Presented at: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 09-12, 2021; Houston, TX p. 1813. [doi: 10.1109/bibm52615.2021.9669477]
28. Ashton R, Leppard B. Differential Diagnosis in Dermatology. In: CRC Press. London: CRC Press; 2014:978-1001.

29. Noy NF, McGuinness DL. Ontology development 101: A guide to creating your first ontology. Corais. 2021. URL: https://corais.org/sites/default/files/ontology_development_101_aguide_to_creating_your_first_ontology.pdf [accessed 2024-05-23]
30. Egenhofer M, Herring J. A mathematical framework for the definition of topological relations. 1990 Presented at: Proceedings of the Fourth International Symposium on Spatial Data Handing; September 1990; California p. 803.
31. Ceusters W. An information artifact ontology perspective on data collections and associated representational artifacts. In: Studies in Health Technology and Informatics. Amsterdam, the Netherlands: IOS Press; 2012:68-72.
32. 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]
33. Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Commun ACM 2016;59(2):44-51. [doi: 10.1145/2844544]
34. 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;74(6):1093-1106 [FREE Full text] [doi: 10.1016/j.jaad.2015.12.038] [Medline: 26896294]
35. Yélamos O, Braun RP, Liopyris K, Wolner ZJ, Kerl K, Gerami P, et al. Dermoscopy and dermatopathology correlates of cutaneous neoplasms. J Am Acad Dermatol 2019;80(2):341-363 [FREE Full text] [doi: 10.1016/j.jaad.2018.07.073] [Medline: 30321581]
36. Draghici C, Vajaitu C, Solomon I, Voiculescu VM, Popa MI, Lupu M. The dermoscopic rainbow pattern—a review of the literature. Acta Dermatovenerol Croat 2019;27(2):111-115. [Medline: 31351506]
37. 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]
38. Burton-Jones A, Storey VC, Sugumaran V, Ahluwalia P. A semiotic metrics suite for assessing the quality of ontologies. Data Knowl Eng 2005;55(1):84-102. [doi: 10.1016/j.datak.2004.11.010]
39. Amith M, Manion F, Liang C, Harris M, Wang D, He Y, et al. Architecture and usability of OntoKeeper, an ontology evaluation tool. BMC Med Inform Decis Mak 2019;19(Suppl 4):152 [FREE Full text] [doi: 10.1186/s12911-019-0859-z] [Medline: 31391056]
40. Obrst L. Ontologies for semantically interoperable systems. 2003 Presented at: Proceedings of the Twelfth International Conference on Information and Knowledge Management; November 3, 2003; New York, NY p. 366-369. [doi: 10.1145/956863.956932]
41. Mullin S, Zola J, Lee R, Hu J, MacKenzie B, Brickman A, et al. Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. J Biomed Inform 2021;122:103889. [doi: 10.1016/j.jbi.2021.103889] [Medline: 34411708]
42. Sahoo SS, Kobow K, Zhang J, Buchhalter J, Dayyani M, Upadhyaya DP, et al. Ontology-based feature engineering in machine learning workflows for heterogeneous epilepsy patient records. Sci Rep 2022;12(1):19430 [FREE Full text] [doi: 10.1038/s41598-022-23101-3] [Medline: 36371527]
43. Zemmouchi-Ghomari L. Ontology and machine learning: a two-way street to improved knowledge representation and algorithm accuracy. : Springer Nature; 2023 Presented at: Proceedings of International Conference on Paradigms of Communication, Computing and Data Analytics; October 11, 2023; Singapore p. 181-189. [doi: 10.1007/978-981-99-4626-6_15]
44. Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/ artificial intelligence for the diagnosis of melanoma. Br J Dermatol 2009;161(3):591-604. [doi: 10.1111/j.1365-2133.2009.09093.x] [Medline: 19302072]
45. Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A, International Skin Imaging Collaboration. Results of the 2016 International Skin Imaging Collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78(2):270-277. [doi: 10.1016/j.jaad.2017.08.016] [Medline: 28969863]
46. Pham TC, Luong CM, Hoang VD, Doucet A. AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. Sci Rep 2021;11(1):17485 [FREE Full text] [doi: 10.1038/s41598-021-96707-8] [Medline: 34471174]
47. Tschandl P, Kittler H, Argenziano G. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. Br J Dermatol 2017;177(3):867-869. [doi: 10.1111/bjd.15695] [Medline: 28569993]
48. Shetty B, Fernandes R, Rodrigues AP, Chengoden R, Bhattacharya S, Lakshmanna K. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. Sci Rep 2022;12(1):18134. [doi: 10.1038/s41598-022-22644-9] [Medline: 36307467]

Abbreviations

CNN: convolutional neural network
D3X: dermoscopy differential diagnosis explorer
DERMLEX: dermatology lexicon
DERMO: human dermatological disease ontology
DEVO: dermoscopy elements of visuals ontology
EVO: elements of visuals ontology
IAO: information artifact ontology
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
LOINC: logical observation identifier names and codes
MedDRA: Medical Dictionary for Regulatory Activities
NCBO: National Center for Biomedical Ontology
NCI: National Cancer Institute
OBO: Open Biological and Biomedical Ontology
OWL: web ontology language
PCP: primary care physician
rdf: resource description framework
RDF: resource description framework
SHACL: shapes constraint language
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms
SPO: skin physiology ontology
SWO: software ontology

Edited by C Lovis, G Eysenbach; submitted 12.07.23; peer-reviewed by C Gaudet-Blavignac; comments to author 09.12.23; revised version received 18.04.24; accepted 04.05.24; published 21.06.24.

Please cite as:
Lin RZ, Amith MT, Wang CX, Strickley J, Tao C
Dermoscopy Differential Diagnosis Explorer (D3X) Ontology to Aggregate and Link Dermoscopic Patterns to Differential Diagnoses: Development and Usability Study
JMIR Med Inform 2024;12:e49613
URL: https://medinform.jmir.org/2024/1/e49613
doi:10.2196/49613
PMID:38904996

Original Paper

An Ontology-Based Decision Support System for Tailored Clinical Nutrition Recommendations for Patients With Chronic Obstructive Pulmonary Disease: Development and Acceptability Study

Daniele Spoladore[1,2*], DPhil; Vera Colombo[1*], DPhil; Alessia Fumagalli[3*], MD, DPhil; Martina Tosi[4,5*], RD; Erna Cecilia Lorenzini[4,6*], MD; Marco Sacco[1*]

[1]Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, National Research Council of Italy, Lecco, Italy

[2]Department of Pure and Applied Sciences, Computer Science Division, Insubria University, Varese, Italy

[3]Unit of Pulmonary Rehabilitation, IRCCS, Italian National Research Center on Aging, Casatenovo, Italy

[4]Institute of Agricultural Biology and Biotechnology, National Research Council of Italy, Milan, Italy

[5]Department of Health Science, University of Milan, Milan, Italy

[6]Department of Biomedical Sciences for Health, Chair of Clinical Pathology, University of Milan, Milan, Italy

[*]all authors contributed equally

Corresponding Author:
Daniele Spoladore, DPhil
Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing
National Research Council of Italy
Via G. Previati 1E
Lecco, 23900
Italy
Phone: 39 03412350202
Email: daniele.spoladore@stiima.cnr.it

Abstract

Background: Chronic obstructive pulmonary disease (COPD) is a chronic condition among the main causes of morbidity and mortality worldwide, representing a burden on health care systems. Scientific literature highlights that nutrition is pivotal in respiratory inflammatory processes connected to COPD, including exacerbations. Patients with COPD have an increased risk of developing nutrition-related comorbidities, such as diabetes, cardiovascular diseases, and malnutrition. Moreover, these patients often manifest sarcopenia and cachexia. Therefore, an adequate nutritional assessment and therapy are essential to help individuals with COPD in managing the progress of the disease. However, the role of nutrition in pulmonary rehabilitation (PR) programs is often underestimated due to a lack of resources and dedicated services, mostly because pneumologists may lack the specialized training for such a discipline.

Objective: This work proposes a novel knowledge-based decision support system to support pneumologists in considering nutritional aspects in PR. The system provides clinicians with patient-tailored dietary recommendations leveraging expert knowledge.

Methods: The expert knowledge—acquired from experts and clinical literature—was formalized in domain ontologies and rules, which were developed leveraging the support of Italian clinicians with expertise in the rehabilitation of patients with COPD. Thus, by following an agile ontology engineering methodology, the relevant formal ontologies were developed to act as a backbone for an application targeted at pneumologists. The recommendations provided by the decision support system were validated by a group of nutrition experts, whereas the acceptability of such an application in the context of PR was evaluated by pneumologists.

Results: A total of 7 dieticians (mean age 46.60, SD 13.35 years) were interviewed to assess their level of agreement with the decision support system's recommendations by evaluating 5 patients' health conditions. The preliminary results indicate that the system performed more than adequately (with an overall average score of 4.23, SD 0.52 out of 5 points), providing meaningful and safe recommendations in compliance with clinical practice. With regard to the acceptability of the system by lung specialists (mean age 44.71, SD 11.94 years), the usefulness and relevance of the proposed solution were extremely positive—the scores on each of the perceived usefulness subscales of the technology acceptance model 3 were 4.86 (SD 0.38) out of 5 points, whereas the score on the intention to use subscale was 4.14 (SD 0.38) out of 5 points.

Conclusions: Although designed for the Italian clinical context, the proposed system can be adapted for any other national clinical context by modifying the domain ontologies, thus providing a multidisciplinary approach to the management of patients with COPD.

(JMIR Med Inform 2024;12:e50980) doi:10.2196/50980

KEYWORDS

ontology-based decision support system; nutritional recommendation; chronic obstructive pulmonary disease; clinical decision support system; pulmonary rehabilitation

Introduction

Background

Chronic obstructive pulmonary disease (COPD) is one of the leading causes of morbidity and mortality worldwide. In Italy, 3.2% of the population in 2019 had a diagnosis of COPD, and such numbers are expected to increase in the next years due to the worsening of risk factors [1]. COPD is characterized by chronic inflammation in the lungs and airflow obstruction. Starting from the respiratory system, COPD has several systemic effects, including skeletal muscle wasting (which limits exercise capacity), increased risk of cardiovascular disease, osteoporosis, depression, and anxiety [2]. Pulmonary rehabilitation (PR) is an evidence-based, nonpharmacological intervention that helps people with COPD improve their health condition and quality of life by increasing exercise capacity and reducing symptoms of dyspnea and fatigue [3]. PR is based on a multidisciplinary approach and includes several components, among which nutritional status evaluation is considered relevant.

However, as highlighted recently by the American Thoracic Society, the scarcity of resources often limits the nutritional component, which, in some settings, can be accessible only outside the PR program [4]. Moreover, even though the influence of nutrition on COPD has been recognized and a relationship between some nutrients and pulmonary functions has been identified, PR programs often do not formally include patient-specific dietary recommendations. Therefore, more studies are required to drive the implementation of tailored nutritional strategies [5].

The Role of Nutrition in Patients With COPD: Evidence

The relationship between COPD and nutrition has been investigated by many authors, who have put together a set of evidence that underlines the role of nutrients in PR and the disease exacerbations. The results of the systematic review and meta-analysis by Collins et al [6] demonstrated that nutritional support, primarily in the form of oral nutritional supplements, improves total energy intake, anthropometric measurements, and grip strength in patients with COPD. This work highlighted how dietary changes supported protein and energy intake in patients with COPD, resulting in weight gain and a moderate increase in muscle strength—which can significantly impact respiratory muscle strength, walking distance, and quality of life. Another systematic review [7] found that nutritional supplementation improves body weight gain in patients with COPD, especially when malnourished, in terms of body composition, respiratory muscle strength, and quality of life.

Inadequate nutritional intake, higher daily energy expenditure, weight loss, and fat-free mass (FFM) depletion affect the functional capacity of patients with COPD. The progression of the disease also contributes to an increased respiratory muscle load, hypoxemia, physical inactivity, and release of inflammation mediators [8]. In addition, Liu et al [9] explored the associations among COPD, diet, and inflammation, identifying a correlation between diet and COPD inflammatory status. By calculating the Diet Inflammatory Index in COPD, they observed that COPD was more prevalent in patients with a worse Diet Inflammatory Index.

A systematic review of factors influencing the risk of developing COPD [10] concluded that the Western diet, characterized by the consumption of processed meat and alcohol and a low intake of fruit and vegetables, is associated with the prevalence of COPD. A cross-sectional study [11] assessed the nutritional status of outpatients with COPD and found a significant association between malnutrition and COPD—most of the patients did not have an adequate daily food intake and showed a lower quality of life, highlighting the need for a more tailored nutritional intervention for managing malnutrition in patients with COPD with cachexia. In COPD, low FFM and sarcopenia are predictive factors of mortality [12]. Considering that literature data demonstrate that the presence of malnutrition and, above all, a low FFM index is associated with an increased risk of mortality, Schols et al [13] identified different nutritional risk profiles based on nutritional phenotypes that appear to be predictors of outcome independent of impaired respiratory function. These nutritional phenotypes included obesity, sarcopenic obesity, sarcopenia, and cachexia and required at least three items for their identification: (1) BMI and body circumference, (2) bioelectrical impedance analysis, and (3) history of unintentional weight loss. A recent study [14] demonstrated that patients with COPD with sarcopenia and sarcopenic obesity had worse muscle strength than patients with healthy body weight, claiming that body composition is associated with physical functions. Patients with COPD with obesity have a higher risk of developing comorbidities such as diabetes and metabolic and cardiovascular diseases in addition to typical COPD symptoms.

Research Challenge: Nutritional Status to Prevent COPD Exacerbation

The evidence reported previously indicates that clinicians and dietitians must evaluate the nutritional status and body composition of patients when giving nutritional advice or diet recommendations to patients with COPD with obesity to foster an adequate intake of macro- and micronutrients and control obesity as well [15].

Malnutrition plays a pivotal role among the challenges that must be faced in treating COPD. The study by Mete et al [16] affirms that malnutrition and risk of malnutrition are very frequent conditions among patients with COPD and, together with a significantly lower BMI, are associated with disease severity, as highlighted by pulmonary function tests. In clinical practice, sarcopenia assessment is not part of the standard of care, implying a loss of personalized treatment that should be essential to improve health outcomes related to the disease [17]. The current literature illustrates that screening for sarcopenia and tailored nutrition intervention in patients with COPD could cost-effectively achieve better health outcomes [18,19]. Functional capacity, respiratory muscle strength, and quality of life can all be improved through nutritional supplementation and a better assessment of nutritional status.

The future nutrition challenges require the identification of specific targets of intervention, considering body composition, nutritional status, and inflammation in addition to the strictly clinical aspects. Nowadays, diet is not always an integral part of COPD therapeutic strategy [20]. Moreover, different studies have highlighted the necessity of new methods to assess malnutrition and evaluate nutritional status in patients with COPD, not considering the BMI alone [21] as it misses important changes in body composition.

This new vision places nutritional intervention as an integral part of COPD management, not only in the advanced and early stages but also to prevent the evolution toward respiratory failure. Creating a decision support system (DSS) that brings together data and knowledge in the nutritional field to identify the different nutritional phenotypes and suggest specific dietary recommendations could be useful to meet these challenges. In particular, this DSS could become a valuable tool for pulmonologists to develop nutritional interventions as an integral part of the COPD therapeutic strategy even in the absence of specific resources.

Objective of This Study

Digital health care applications are already used to improve PR models, mainly for self-management; modification of lifestyle factors; and modification of risk factors, such as smoking cessation and fostering physical activity [22]. However, to the best of our knowledge, there are only a few applications focused on nutritional aspects associated with COPD. This work leveraged clinical expert knowledge—formalized into domain ontologies—from the Italian health care context to develop a DSS to support pneumologists in considering nutritional aspects in PR to avoid the disease's exacerbation and provide patients with tailored dietary guidelines. The application exploiting the DSS fits the context of a preprototype according to the World Health Organization (WHO) guidelines for the monitoring and evaluation of digital health interventions [23].

Related Work

Overview

Ontology-based technologies have been adopted for both clinical DSSs and patient-centered DSSs. Regarding the ontological formalization of requisites and tailored suggestions for patients with COPD in DSSs, no work has tackled this issue in PR—except for an early version of the system we propose in this work [24]. Nonetheless, the use of ontologies to formalize COPD has been established in some works. Moreover, specific digital applications for patients with COPD have been developed in the past years. This section presents works relevant to the fields of ontology-based clinical and patient-centered DSSs and digital applications specifically developed for patients with COPD.

Ontologies and Ontology-Based Systems for COPD Management

COPD has been formalized in ontologies or knowledge bases since the early 2010s. The COPD ontology by Greenberg et al [25] was developed to support COPD longitudinal research and clinical trials, leveraging a preexisting model dedicated to representing subpopulations of patients. Cano et al [26] developed a knowledge base devoted to collecting clinical experimental data; the COPD Knowledge Base can semantically map data to physiological and molecular data to support clinical decision-making through a predictive mathematical model.

Ontologies can play a pivotal role in diagnostic systems. In the case of COPD, Rayner et al [27] developed an ontology for the early diagnosis of the illness. The ontological model takes advantage of clinical tests and patients' data (eg, spirometry, forced expiratory volume in the first second [FEV1], age, and smoking status) and tests its robustness against a large data set of patients. The model proved able to categorize patients as "Unlikely COPD," "Probable COPD," and "Definite COPD" cases.

As ontology-enabled reasoning processes are appreciated in the context of clinical decision-making, the adoption of formal models in health care systems aimed at managing chronic conditions—including systems for patients with COPD—has also been established. CHRONIOUS [28] is an open and ubiquitous system that exploits ontology-based inferential reasoning to adapt the platform's services, including monitoring patients' conditions—which is also performed leveraging wearable sensors. Lasierra et al [29] developed an ontology-based telemonitoring system aimed at monitoring patients with chronic diseases at home. The ontology layer (Home Ontology for Integrated Management in Home-Based Scenarios) represents the patient profile, vital signs, and chronic conditions (including COPD) to observe the patient's evolution and plan activities. Similarly, Ajami and McHeick [30] developed a domain ontology encompassing environmental features, patient data, clinical status and diseases (including COPD), and devices to monitor and identify patients' conditions. This DSS aimed to foster patients' adherence to a healthy lifestyle and identify and avoid possibly dangerous situations.

Digital Applications for Patients With COPD

Digital applications may offer solutions to both clinicians—to ease the process of assessment and support clinical decisions—and patients—mainly to support them in the management of the disease. Digital applications currently available for patients with COPD are mainly telemedicine and telerehabilitation solutions. Most of them are focused on educational programs, symptom tracking, behavior change,

support for medication or treatment, and activity report [31]. In some cases, as both research prototypes [32] and commercial products [33], the educational programs include specific sections on nutrition, in which helpful tips and recommendations are provided to manage symptoms during meal consumption and help maintain a balanced diet. However, such applications do not provide personalized information tailored to the patient's nutritional status. In such cases, the applications targeted at health care professionals comprise the "clinician side" of telemedicine platforms (ie, monitoring dashboard allowing for remote visualization of patients' data and applications for remote teleconsulting).

Differently from the aforementioned solutions, the DSS proposed in this work tackles the role of nutritional therapy in the PR of patients with COPD—an aspect neglected in existing DSSs.

Methods

A DSS was developed to tackle the aims described in the Introduction section and the research challenge described in the Objective of This Study section, thus supporting pneumologists in suggesting specific dietary recommendations for patients with COPD. The DSS leverages expert knowledge in the form of ontologies and, through automatic inference processes, is expected to provide guidelines to prepare a tailored dietary plan.

The development of the ontology underlying the DSS leveraged expert clinical knowledge to maximize its acceptability. Contrary to purely data-driven approaches, ontologies formalize information to enable a system to perform inferences (based on the knowledge formalized in the ontology). The inference process simulates human inference capabilities [34] so that ontology-based approaches are perceived as transparent. Thus, ontologies are widely adopted in several artificial intelligence and health-related applications [35].

However, the development of a domain ontology is not a trivial task—it is a process that may involve several activities (eg, the acquisition of knowledge, its conceptualization, the survey of existing models that can be reused, the development of the model in a formal language, and the testing of the developed ontology [36]). The ontology engineered for the proposed DSS was developed following the Agile, Simplified, and Collaborative Ontology Engineering Methodology (AgiSCOnt) [37] engineering methodology, which involves knowledge elicitation techniques and domain experts in the development phase of the ontology. This collaborative ontology engineering methodology adopts unstructured interviews, scientific literature surveys, and discussions to elicit the necessary knowledge to minimize the impact of the "knowledge elicitation bottleneck" (ie, it takes longer to gather knowledge from experts and documentation than to write the software [38]). Its collaborative and agile features and validation [37] were the reasons behind the adoption of AgiSCOnt in this work.

The methodology involves three phases:

1. *Domain analysis and conceptualization*, which includes the identification of the knowledge to be included and the preparation of competency questions (CQs) [39] that the ontology is expected to answer; it enables the conceptualization of the domain (which results in a conceptual map) and the preliminary identification of existing ontological resources that can be reused.
2. *Development and testing*, which involves the selection of the ontological languages to formalize the conceptualization developed in the previous step and the identification of ontology design patterns (ie, "micro-ontologies" that can be reused to model recurrent problems in ontology engineering [40]). This step results in the prototypization of the ontology, which undergoes a preliminary test to assess the validity of its inferences.
3. *Ontology use and updating*, which includes activities such as use of the developed ontology in an application, extended validation, and feedback collection. The following subsections delve into the engineering process.

The development of this ontology involved the following team members: 1 ontologist with experience in modeling using agile ontology engineering methodologies, 1 biomedical engineer with previous experience in ontology engineering and knowledge of COPD, 2 senior dieticians with clinical experience with patients with COPD, and 1 senior pneumologist. The team was composed of clinical personnel from universities (dieticians) and a research and cure center (Scientific Institute for Research, Hospitalisation and Health Care) with a specialization in COPD (pneumologist) and yearly experience treating such patients. The team delved into nutrition and diet's role in tackling this disease, with examples from the literature and clinical trials in which the clinical personnel was involved. The discussion was then oriented to identify some of the issues that the ontology was expected to answer (ie, the CQs).

Ethical Considerations

This study does not include human subjects research (no human subjects experimentation or intervention was conducted) and so does not require institutional review board approval.

Results

This section describes the development of the COPD and Nutrition domain ontology for the DSS and the ontology-based application for clinical personnel.

The COPD and Nutrition Domain Ontology

Domain Analysis and Conceptualization

The considerations reported in The Role of Nutrition in Patients With COPD: Evidence and Research Challenge: Nutritional Status to Prevent COPD Exacerbation sections were gathered by the team in this phase. Leveraging the objective (ie, providing clinical personnel with support in identifying tailored nutritional recommendations for patients with COPD), the team decided that the ontology should focus on representing the patients' health condition and the stage of their COPD. On the basis of their expertise, clinicians pointed out that the purpose of the ontology should be to illustrate, for each patient, a tailored percentage of macro- and micronutrients they are advised to consume on a daily basis to avoid exacerbations, as well as to provide nutritional guidance. The summary of the entities and

expected outputs of the ontology is provided in Textbox 1, listing the CQs and their answers.

Textbox 1. The list of competency questions and answers for the Chronic Obstructive Pulmonary Disease (COPD) and Nutrition ontology engineering process.

What information identifies a patient? What basic information is used to identify the patient? What clinical information is used to identify the patient?

- A patient is identified by an ID and gender. Each patient is associated with 1 health condition and 1 anthropometric phenotype (defined via BMI cutoffs). Each patient can be classified as a patient with sarcopenia or cachexia or as a patient without sarcopenia or cachexia.

How is COPD evaluated?

- COPD is evaluated according to the criteria defined in the gold standard—it can be mild, moderate, severe, and very severe. The criterion to be analyzed is the forced expiratory volume in the first second.

How is sarcopenia evaluated?

- The status of sarcopenia is evaluated according to clinical standards (operational definition of sarcopenia). The first criterion comprises low muscle strength, the second criterion comprises a low muscle quantity or quality, and the third criterion comprises low physical performance. The presence of the first criterion alone indicates probable sarcopenia, the presence of both the first and second criteria indicates diagnosed sarcopenia, and the presence of all 3 criteria indicates severe sarcopenia.

How is cachexia evaluated?

- Cachexia is evaluated by means of biochemical indicators according to the study by Evans et al [41]. Albuminemia, iron transport, and polymerase chain reaction (PCR) criteria must be copresent to indicate a cachexia diagnosis.

Which data characterize the patient's health condition? How is the nutritional risk index assessed?

- Patients' health condition must indicate the stage of COPD and the nutritional risk index profile characterizing the patient. Each health condition must illustrate anthropometric measures (current weight, usual weight, height in meters, and BMI), physical performance indicators (hand grip and gait speed), and biochemical indicators (albuminemia, PCR, resistance, reactance, and iron transport). The nutritional risk index assessment is performed following clinical standards.

What recommendations are given to clinical personnel?

- The recommendations provided to clinical personnel indicate (for each patient) the basal metabolic rate and the corrected caloric intake, the daily macronutrient shares (protein, minimum and maximum share of carbohydrates, minimum and maximum share of fats, minimum and maximum share of fiber, maximum share of sugar, and maximum share of saturated fats), the amount of cholesterol and sodium, and whether the patient should increase their caloric intake by means of branched-chain amino acid or energy-protein supplementations.

How are recommendation values calculated?

- The indications provided in the patient's recommendation are calculated according to clinical standards and differentiated according to the patient's gender, stage of COPD, and anthropometric phenotype.

In this phase, the pneumologist specialized in clinical nutrition and a clinical dietitian defined the phenotypes and their nutritional requirements based on anthropometric and clinical parameters and comorbidities according to national and international guidelines. For the definition of each anthropometric phenotype, COPD stage (defined via FEV1 and forced vital capacity and in particular FEV1–to–forced vital capacity ratio [42,43]; Table 1) and the presence of sarcopenia or cachexia were considered. A total of 5 metabolic phenotypes for patients with COPD were developed (underweight, normal weight, overweight, first-degree obesity, and second-degree obesity), and each can be characterized by the presence of sarcopenia, cachexia, or none of the 2 conditions (Table 2).

Moreover, nutritional risk was assessed by means of the nutritional risk index (NRI) formula [44] as follows:

(1)

Thus, patients were classified according to this formula (Table 3).

The diagnosis of sarcopenia was based on the analysis of specific patients' value indicators. The first one was the appendicular skeletal muscular mass, which is calculated according to the work by Sergi et al [45]:

(2)

Appendicular skeletal muscular mass and other indicators allow for the classification of patients' sarcopenic condition according to the 3 criteria in Table 4 [46].

If the first criterion applied, then the patient was *probable* sarcopenic; if the first and second criteria applied, then the patient was *diagnosed* sarcopenic; if all 3 criteria applied, then the patient was *severe* sarcopenic.

A similar approach was adopted to identify whether a patient was cachectic. If a patient's polymerase chain reaction was >10, the level of iron transport was <150, the level of albuminemia was <3.5, and the patient was sarcopenic, then they were also cachectic [41]. As such, it is safe to infer that cachexia is a particular case of sarcopenia.

Calculation of nutritional recommendations was performed using the "if-then" type rules produced by clinical personnel considering reference values reported in scientific literature and national and international guidelines. The COPD DSS provides the following nutritional information and recommendations: (1) basal metabolic rate (BMR), (2) total daily energy requirement (kcal), (3) meal frequency, (4) indication for energy-protein supplementation (yes or no), (5) indication for branched-chain amino acid (BCAA) supplementation (yes or no), (6) protein intake (percentage and grams), (7) carbohydrate intake (percentage and grams), (8) lipid intake (percentage and grams), (9) sugar intake (maximum percentage and grams), (10) saturated fat intake (maximum percentage and grams), (11) cholesterol intake (maximum milligrams), (12) fiber intake (minimum and maximum grams), (13) sodium intake (maximum milligrams), and (14) calcium intake (milligrams).

For BMR calculation, Harris-Benedict or Mifflin predictive equations based on sex, age, weight, and height were considered. The Harris-Benedict equation was preferred for patients who were underweight or had normal weight, whereas the Mifflin equation was used for patients with COPD with overweight or obesity [47-49]. According to the COPD stage, different correction factors to BMR were applied (Multimedia Appendix 1)—for patients who were underweight, had a normal weight, were overweight, or had first- or second-class obesity presenting with COPD at the first or second stage and who were nonsarcopenic and noncachectic or sarcopenic, a 1.5 correction factor was used to calculate total daily energy requirement; for the same categories of patients presenting with COPD at the third or fourth stage, a correction factor of 1.8 was preferred to counteract the important energy expenditure due to the respiratory work of these patients. Only for patients with cachexia a 1.8 correction factor was always applied regardless of BMI and COPD stage.

The impact of physical activity was deemed marginal for the correction factor's definition as patients with COPD are a vulnerable population presenting with comorbidities that limit their capacity to perform physical activity. As seen in the pathological lung mechanics of patients with COPD, dynamic hyperinflation influences the proper operation of the chest's horizontal and vertical diameters that expand during inspiration due to the activation of the external intercostal muscles and the diaphragmatic contraction. This impairment plays a role in how much exercise a patient with COPD can tolerate. As shown in individuals with severe-stage COPD and weight loss related to COPD, respiratory muscle weakness exacerbates the breathing mechanism. Increased dyspnea and decreased exercise tolerance are directly related to this respiratory muscle weakening. Moreover, it has been observed that patients with COPD are characterized by higher levels of physical inactivity [50]. As far as protein requirement was concerned, clinical experts decided to provide different recommendations according to real or ideal weight or BMI or focusing on FFM considering specific body composition or COPD stage to give personalized recommendations to not compromise health and nutritional status as well as prevent further decrease in metabolically active lean mass [12]. For this reason, cachectic phenotype, regardless of BMI and COPD stage, was always given a high percentage of protein intake. BCAA was suggested when energy requirements but not protein requirements were met through diet. Differently, protein-caloric supplements were indicated when neither energy nor protein requirements were satisfied through food intake [51]. Phenotypes characterized by cachexia, regardless of BMI, underweight sarcopenic and normal-weight phenotype, and COPD stage, were always recommended BCAA and energy-protein supplementation in consideration of their health status. For the same reason, BCAA supplementation was always suggested for sarcopenic phenotypes [52]. As far as carbohydrate metabolism was concerned, and in line with the lower levels reported in the reference values of nutrients and energy for the Italian population [13,47], a carbohydrate intake of 45% to 50% of total daily calories and a maximum sugar intake of 15% of total calories were indicated except for individuals with type 2 diabetes mellitus, for whom a maximum of 10% of total calories was indicated to be reached through sugar intake. A lower recommendation for carbohydrate intake permits the promotion of protein and lipid intake, which are functional for respiratory work and to prevent further weight or lean mass loss. Regarding dietary fiber, it was decided that providing a lower indication than that provided by the Livelli di Assunzione di Riferimento di Nutrienti ed energia (LARN; National Recommended Energy and Nutrient Intake Levels) was preferred to encourage the intake of energy and protein foods, especially considering the difficulties met by patients with COPD in feeding and the early sense of satiety as a result of high-fiber food intake. Compared to the LARN, a higher intake of lipids was recommended, especially for phenotypes presenting with a partial pressure of carbon dioxide of >50 mm Hg, a measure of carbon dioxide in arterial or venous blood [53]. For saturated fats, a maximum intake of 10% of total daily calories was suggested according to LARN guidelines. Regarding cholesterol intake, a LARN nutritional goal for prevention, a maximum of 300 mg per day was recommended except for patients with high cholesterol levels, for whom the target was lowered to 200 mg per day. Hypercholesterolemia was defined as elevated total or low-density lipoprotein cholesterol levels or low levels of high-density lipoprotein cholesterol. Regarding micronutrients, sodium and calcium intake was considered. According to the recent WHO report on sodium intake [54], a maximum intake of <2000 mg per day of sodium (<5 g per day of salt) was recommended. A slightly higher recommendation than that of the LARN was given for calcium to prevent or counteract osteoporosis [55]. Finally, given the difficulties in feeding and early satiety observed, an indication for fractioned meals was given to meet energy and nutritional requirements throughout the day with small and frequent meals.

All the parameters involved in the evaluations presented previously (ie, FEV1; partial pressure of carbon dioxide; resistance; reactance; iron transport; albuminemia; polymerase chain reaction; hand grip; gait speed; and general patient

information such as age, gender, height in meters, current weight, and usual weight) are usually acquired during patient assessment (such as spirometry and blood tests). AgiSCOnt's outputs for the domain analysis phase consisted of the conceptual map reported in Figure 1 and a list of CQs (Textbox 1).

Table 1. The cutoffs identifying the chronic obstructive pulmonary disease (COPD) stages based on the forced expiratory volume in the first second (FEV1) values.

COPD stage number	COPD stage name	FEV1 (%)
I	Mild	≥80
II	Moderate	≥50 to <80
III	Severe	≥30 to <50
IV	Very severe	<30

Table 2. World Health Organization cutoffs for nutritional status categories based on BMI.

Nutritional status	BMI (kg/m^2)
Underweight	<18.5
Healthy weight	≥18.5 to ≤24.9
Overweight (preobesity)	≥25.0 to <29.9
Obesity degree I	≥30.0 to <34.9
Obesity degree II	≥35.0 to <39.9
Obesity degree III	≥40

Table 3. The cutoffs identifying the 4 levels of nutritional risk based on the nutritional risk index (NRI).

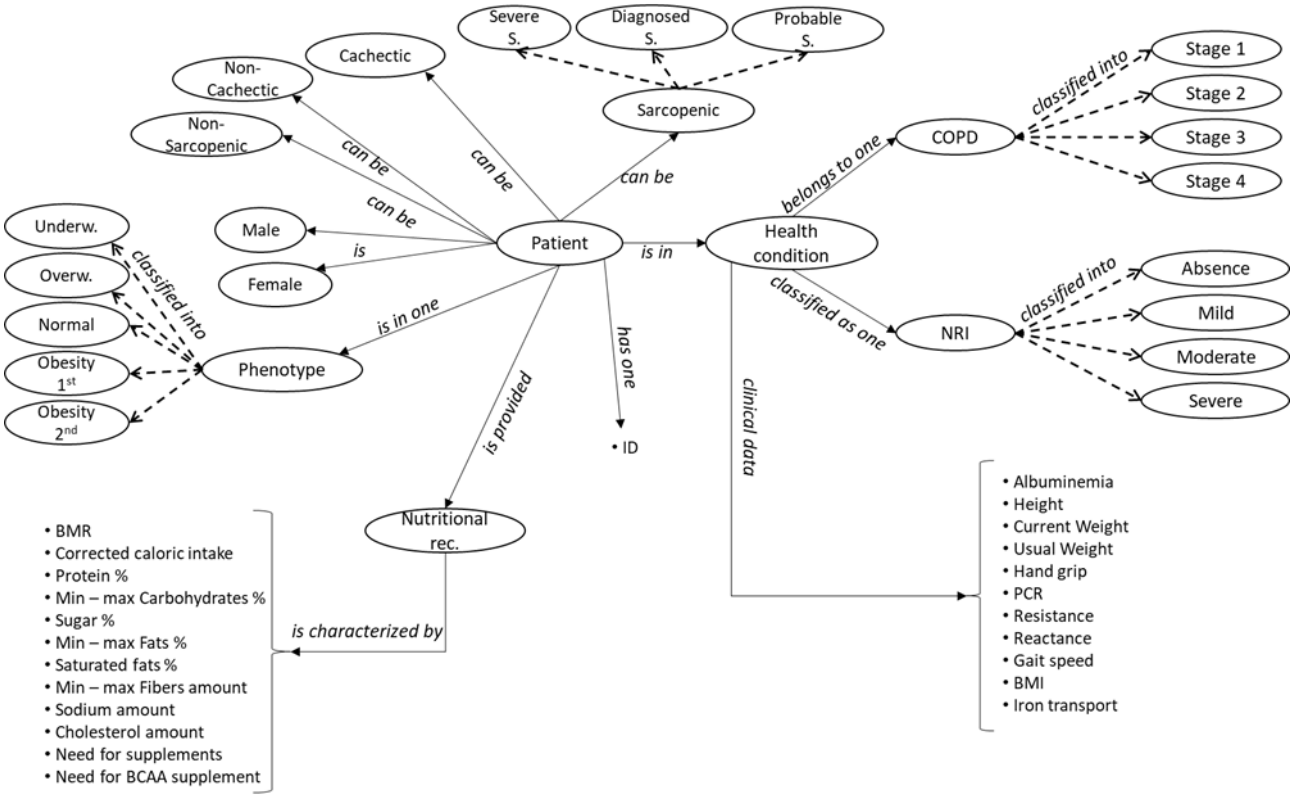
NRI	Index
Absence of risk	>100
Mild risk	≥97.5 to ≤100
Moderate risk	≥83.5 to <97.5
Severe risk	<83.5

Table 4. Criteria for the classification of sarcopenia in patients.

Criteria	Cutoff for male patients	Cutoff for female patients
Low muscular strength—hand grip	<27 kg	<16 kg
Low muscular quantity	ASMM[a]<20 kg; ASMM/height^2<7 kg/m^2	ASMM<15 kg; ASMM/height^2<5.5 kg/m^2
Poor physical performance	≤0.8 m/s	≤0.8 m/s

[a]ASMM: appendicular skeletal muscular mass.

Figure 1. An excerpt of the conceptual map developed by the team involved in the ontology engineering process for the clinical chronic obstructive pulmonary disease (COPD) and nutrition ontology. BCAA: branched-chain amino acid; BMR: basal metabolic rate; NRI: nutritional risk index; PCR: polymerase chain reaction.



Development and Testing

The development phase adopted the conceptual map and CQs produced in the previous phase to guide the entire development process and discuss whether to model some concepts pertaining to the patients or their health conditions. The ontology editor Protégé (Stanford Center for Biomedical Informatics Research) [56] that supports Resource Description Framework [57] and Web Ontology Language [58] with a DL (description logic) profile [59] was adopted. Clinicians explicitly asked any significant advancement in the development of TBox and Abox (eg, patient modeling, health condition characterization, and recommendation modeling) to be illustrated to ensure that undesired entailments were not modeled in the ontology. From a reuse perspective, no ontology able to describe the conceptualization reached in this study was found; the only ontology design pattern reused in this ontology was the one that relates a copd:Patient to their copd:Health_Condition via the copd:isInHealthCondition object property [60]. The developed ontology (prefixed with copd:), discussed in this subsection, is accessible in Multimedia Appendix 2.

The development started with the identification of concepts that could be translated into owl:Classes. The concept of copd:Patient is pivotal in this ontology. Each patient is defined by exactly 1 patient ID, is given at least 1 nutritional recommendation individual, and is characterized by a health condition individual.

Each patient needs to be classified as copd:Female or copd:Male—disjoint classes—and as copd:Cachectic (or its complement copd:non-Cachectic) or copd:Sarcopenic (or its complement copd:non-Sarcopenic). Sarcopenia and cachexia are modeled as attributes of the patient and not of their health condition. The clinical personnel deemed essential to state that these 2 conditions have systemic status; therefore, they characterize the individual as a whole. The class copd:Sarcopenic is further detailed into the subclasses copd:Probable_Sarcopenic, copd:Diagnosed_Sarcopenic, and copd:Severe_Sarcopenic to reflect the operational definition standard provided by clinicians (presented in the previous subsection).

In the same way, the copd:Anthropometric_Phenotypes are characteristics of the copd:Patients, and this class lists 5 subclasses for the representation of the phenotypes.

Similarly to copd:Patient, the development of the TBox pertaining to the patient's health condition was discussed among the team members—each health condition is characterized by an NRI profile, but in general, a copd:Health_Condition is not necessarily characterized by COPD. The terms adopted in the conceptual map to sketch the relationships among copd:Health_Condition, copd:Nutritional_Risk_Index_Profile, and copd:COPD_HC were found indicative of the clinicians' perspective—both NRI and COPD are considered particular attributes of a health condition (ie, there could be health conditions characterized only by an NRI profile but lacking COPD). Therefore, the classes copd:Nutritional_Risk_Index_Profile and copd:COPD_HC were modeled as rdfs:subclassOf copd:Health_Condition. This decision was also encouraged by the fact that the datatype properties copd:FEV1 and copd:nutritionalRiskIndex have copd:Health_Condition as the domain.

The copd:Nutritional_Risk_Index_Profile and copd:COPD_HC subclasses are characterized by restrictions that allow for the classification of individual health conditions whose copd:nutritionalRiskIndex and copd:FEV1 object values fall under specific restrictions.

Each copd:Health_Condition is described by a set of datatype properties, which represent the clinical data elicited in the previous phase and are required to enable the patient's classification and recommendations. Each owl:Individual belonging to this class also materializes inferred triples related to the copd:AppendicularSkeletalMuscleMass, the copd:ResistiveIndex, and the copd:nutritionRiskIndex. While the copd:nutritionalRiskIndex is calculated using semantic web rule language (SWRL) rules and used to classify each copd:Health_Condition into one of the NRI's subclasses, the copd:ResistiveIndex is a piece of information necessary to calculate the copd:AppendicularSkeletalMuscleMass (both are inferred as the result of 2 different SWRL rules). Figure 2 illustrates an example of copd:Health_Condition completed with all its datatype properties (both asserted and inferred).

The ontology makes use of 39 datatype properties and 2 object properties (copd:isInHealthCondition and copd:hasRecommendation)—almost all datatype properties were used to provide values for the patient's health condition and nutritional recommendation.

The ontology also contains 79 SWRL rules, which are largely used to represent the tuples in Multimedia Appendix 1, depicting the conditions that determine the shares and amounts of nutrients characterizing a patient's diet (see the full ontology in Multimedia Appendix 2). The equations adopted to calculate the BMR and the corrected caloric intake were adapted with SWRL using mathematical built-ins [61]. Taking as an example a copd:Overweight, copd:non-Sarcopenic, and copd:non-Cachectic male patient characterized by copd:Stage2 disease, the BMR is inferred through the following rule (for each male patient not characterized by sarcopenia or cachexia, calculate the BMR using the Harris-Benedict equation and round the result):

> *Male(?p), Overweight(?p), (not (Cachectic))(?p), (not (Sarcopenic))(?p), hasRecommendation(?p, ?rec), isInHealthCondition(?p, ?hc), age(?hc, ?age), currentWeight(?hc, ?kg), height_meters(?hc, ?m), multiply(?a, ?kg, 13.75), multiply(?b, ?m, 5, ?100), multiply(?c, 6.78, ?age), add(?d, 66.5, ?a, ?b), subtract (?e, ?d, ?c), round(?f, ?e) -> regularRecommendedCaloricIntake(?rec, ?f)*

Then, the correction is applied (for each male patient not characterized by sarcopenia or cachexia with stage-1 or stage-2 COPD, correct the BMR calculated using the Harris-Benedict equation by multiplying it by 1.5):

> *(Normal_Weight or Obesity_1st_Degree or Overweight or Underweight)(?p), hasRecommendation(?p, ?rec), isInHealthCondition(?p, ?hc), (Stage1 or Stage2)(?hc), (not (Cachectic))(?p), (not (Sarcopenic))(?p), regularRecommendedCaloricIntake(?rec, ?reg), multiply(?corin, ?reg, 1.5), round(?f, ?corin) -> correctedRecommendedCaloricIntake(?rec, ?f)*

The definition of the share of protein that a patient with COPD needs is calculated by identifying the amount (in grams) of protein. With the sole exception of patients with cachexia—who are given a 25% protein share for clinical reasons—for each patient, the daily quantity of protein is calculated according to their weight (for each patient with normal or overweight status and not characterized by sarcopenia or cachexia, obtain the amount of protein in grams by multiplying their current weight by 1.2):

> *(Normal_Weight or Overweight)(?p), (not (Cachectic))(?p), (not (Sarcopenic))(?p), hasRecommendation(?p, ?rec), isInHealthCondition (?p, ?hc), currentWeight (?hc, ?w), multiply (?pgra, ?w, 1.2) -> proteinsGrams (?rec, ?pgra)*

This amount is then converted into calories, bearing in mind that 1 protein is equal to 4 kcal, and then the daily protein share is calculated. This approach also enables the possibility to correct the amount of protein for particular classes of patients; for example, dieticians indicated that patients classified as copd:non-Sarcopenic, copd:non-Cachectic, and copd:Underweight should have their protein share calculated considering a different BMI (which is set to a higher value to fight their underweight condition and is established at 22.5 kg/m^2).

As mentioned previously, the ontology provides enough SWRL rules to model all the information identified by the domain experts and elicited in Multimedia Appendix 1. As established by the development step in AgiSCOnt, the ontology underwent a test with data from 6 patients provided by clinicians. The test was divided into 2 steps. The first was dedicated to assessing whether the ontology provided a correct classification of the patients (ie, whether it identified copd:Sarcopenic and copd:Cachectic status for each patient and whether the stage of COPD and the copd:Nutritional_Risk_Index_Profile were correctly inferred). Thus, by querying the ontology using SPARQL (World Wide Web Consortium) [62], it was possible to assess the accuracy of the inferences (reported in Table 5).

The pneumologist and dieticians verified the correctness of the classification for each patient. All 6 individuals representing patients were found to be correctly classified. The second phase dealt with the retrieval of nutritional suggestions and their evaluation by the clinical personnel with the aim of assessing the validity of the SWRL rules modeled in the COPD and Nutrition ontology. The ontology was queried using SPARQL to retrieve all the nutrient minimum and maximum shares and quantities deemed important for patients with COPD (the results of the query are reported in Multimedia Appendix 3). Each copd:Nutritional_Recommendation and its inferred nutrient values were evaluated by clinical personnel and were found to be correct—although, for some values such as the copd:proteinShare and copd:fiberMINamount, a rounding of the decimal was suggested by dieticians.

Figure 2. The complete datatype property set for a copd:Health_Condition. The datatype properties with a yellow background represent inferred values.

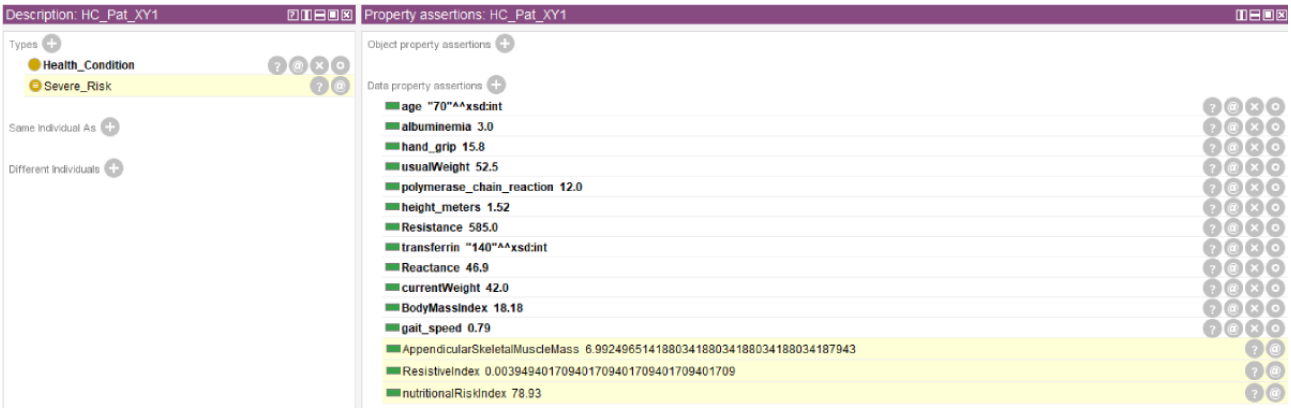


Table 5. An excerpt of the results retrieved for the query investigating, for each patient, their ID, their status (whether they had sarcopenia or cachexia), the stage of chronic obstructive pulmonary disease (COPD), and their nutritional risk index profile. For patients characterized by sarcopenia, all the subclasses of copd:Sarcopenic are illustrated so that clinical personnel can easily see the importance of this condition.

?id	?status	?antrPhen	?copdstage	?nri
001	copd:Cachectic	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Diagnosed_Sarcopenic	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Female	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Probable_Sarcopenic	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Sarcopenic	copd:Underweight	copd:Stage1	copd:Severe_Risk
001	copd:Severe_Sarcopenic	copd:Underweight	copd:Stage1	copd:Severe_Risk
...[a]	...	...	...	...
BB	copd:Female	copd:Normal_Weight	copd:Stage4	copd:Absence_of_Risk
BB	copd:non-Cachetic	copd:Normal_Weight	copd:Stage4	copd:Absence_of_Risk
BB	copd:non-Sarcopenic	copd:Normal_Weight	copd:Stage4	copd:Absence_of_Risk
CV	copd:Male	copd:Obesity_1st_Degree	copd:Stage3	copd:Mild_Risk
CV	copd:non-Cachetic	copd:Obesity_1st_Degree	copd:Stage3	copd:Mild_Risk
CV	copd:non-Sarcopenic	copd:Obesity_1st_Degree	copd:Stage3	copd:Mild_Risk
...	...	...	...	...

[a]Data not reported for conciseness.

Ontology Use and Updating

The COPD and Nutrition ontology described in the previous subsection is a tested prototype able to classify patients properly and provide clinicians with nutrition-related recommendations. To support clinical personnel, the ontology-based DSS needs to be integrated into digital applications, enabling clinicians to access its functionalities intuitively and easily (as described in the following section). This would also enable the assessment of the usefulness and accuracy of the inferences by leveraging on clinicians outside the development team.

The Application for Clinical Personnel

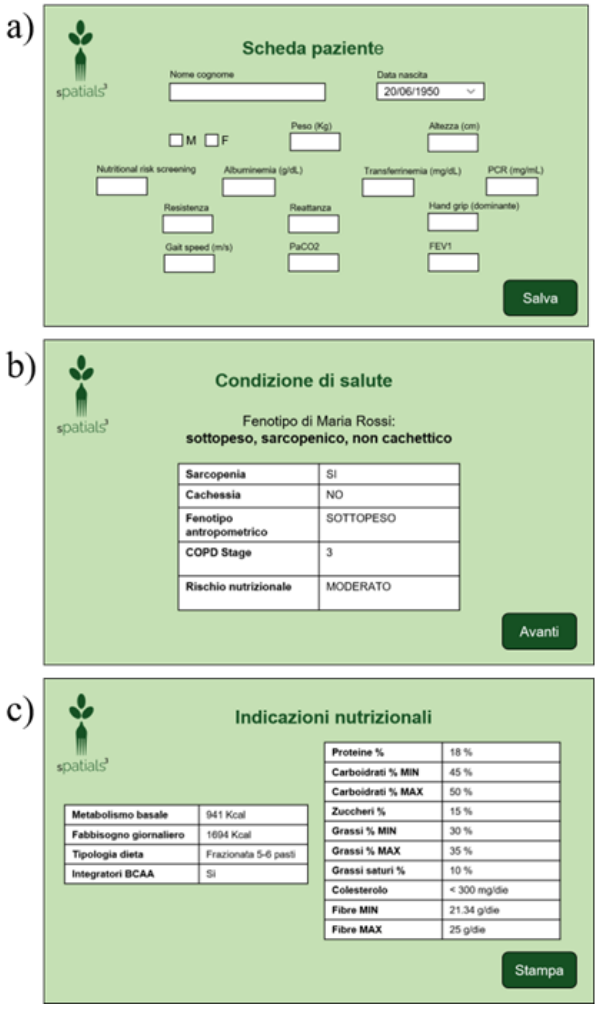
The clinician application was developed to run on a Windows PC or laptop and allows the lung specialist to obtain an overview of the patient's nutritional condition and generate tailored dietary recommendations. To achieve this, the application is connected to the DSS (hosted on a semantic repository) with permission to modify the ontology, reason over new input data, and obtain a new set of dietary recommendations. The application and DSS communication are based on SPARQL queries running over the Stardog reasoner. The entire architecture has already been tested in previous work [24,63].

The application is based on a simple and intuitive graphical user interface (GUI). Its flow is as follows:

1. The clinician logs in to the application either to create a new patient profile or to modify an existing one. The Patient Profile (Scheda Paziente) panel (Figure 3A) has several fields corresponding to the input information needed by the DSS to represent the user's health condition. These include personal data and the results of the patient assessment.
2. Once the new patient profile is saved, the application sends a SPARQL query INSERT to upload the new information into the ontology. The DSS reasons over the new data to obtain the patient's classification and the nutritional recommendations. The output is sent back to the application in the form of a JSON file to populate the GUI panels.

3. The Health Condition (Condizione di Salute) panel, shown in Figure 3B, shows the classification results: metabolic phenotype, presence of sarcopenia and cachexia, anthropometric phenotype, COPD stage, and NRI.
4. The subsequent panel, the Nutritional Recommendations (Indicazioni Nutrizionali), shown in Figure 3C, summarizes the nutritional recommendations, indicating basal metabolism, daily intake, type of diet, suggested BCAA supplement, percentage, and grams of micro- and macronutrients.
5. The clinician can save the recommendations in a PDF file or print them to share them with the dietician to support them in preparing a diet for the patient.

Figure 3. The 3 main graphical user interface panels of the clinician application: (A) patient profile, (B) health condition, and (C) nutritional recommendations.



Preliminary Validation With Clinical Personnel

Overview

Before the implementation in a real use-case scenario, we performed an expert validation of the DSS as a preliminary but necessary step. In total, 2 experiments were performed—considering the multidisciplinary nature of the tool, it was necessary to validate 2 aspects of the DSS with 2 different groups of clinicians.

Therefore, each experiment was carried out by a specific group of clinical experts: (1) a group of nutrition experts validated the recommendations generated by the DSS, and (2) a group of specialists in respiratory diseases evaluated the acceptability of the digital application in clinical practice based on the COPD DSS.

Procedure

Participants were recruited through email invitations among the national experts in nutrition and pulmonology, identified by searching the literature and professional networks. Once they expressed willingness to participate in the study, they signed a written informed consent form and agreed to the data treatment according to the General Data Protection Regulation. In total, 2 experimenters scheduled the web-based video calls—one for each participant—between November 2022 and January 2023. During the test, the experimenters briefly introduced the aim of the system and the main steps of the validation and recorded the participants' answers to brief ad hoc questionnaires and spontaneous comments. Descriptive statistics were calculated for each variable, and the spontaneous comments were analyzed and categorized based on their content.

Experiment 1: Validation of the DSS Recommendations

The first experiment focused on validating the nutritional recommendations generated by the DSS. A total of 7 nutrition experts participated in the validation. The experimenters showed the participants the profiles of 5 real patients and the inferred recommendations. The patient profiles were obtained from real clinical cases provided by the clinicians involved in the project. Each profile included the patient ID, age, gender, and health condition containing all the clinical parameters needed as input to the system.

After presenting the patient's condition, the experimenter showed the recommendations generated by the DSS, which included the patient's inferred classification (metabolic phenotype, presence of sarcopenia or cachexia, anthropometric phenotype, COPD stage, and NRI) and the nutritional recommendations with information on metabolism and suggested quantities of macro- and micronutrients. An example of a patient profile used during the experiment is presented in Figure 4, whereas all the patient profiles used during the evaluation are available in Multimedia Appendix 3. Participants were granted up to 15 minutes to observe the presented patient's health condition. During this time, participants were free to perform calculations using the data shown, ask questions (if necessary) to the experimenter, and consult external sources (eg, books and papers). After the 15-minute period, for each patient, we asked the participants to rate on a scale from 1 to 5 how much they agreed with the recommendation; in case of a score of <5, we asked the participant to provide a brief explanation. We also collected spontaneous comments that emerged during the experiment. The maximum duration of the experiment for each participant was 1 hour and 15 minutes.

Figure 4. An example of a patient profile provided to clinicians (on the left) and the inferences drawn by the ontology-based decision support system (on the right).

Records	
ID	BB
Age	85
Gender	F

Clinical data	
Height (m)	1,5
Current weight (Kg)	43
Usual weight (Kg)	43
BMI	19,11

PaCO2 (Hgmm)	40,4
FEV1 (%)	100

CRP (mg/ml)	0,1
Albuminemia (g/dl)	3,86
Transferrin (mg/dl)	264
Resistance	431
Reactance	17,3
Hand grip (dominant) (kg)	16
Gait speed (m/s)	1,31

General inferences	BB
Appendicular Skeletal Muscle Mass	5,193
Resistive Index	0,005
Nutritional Risk Index	100,333
Anthropometric phenotype	Normal weight
Sarcopenia	no
Cachexia	no
Nutritional risk	Absent
COPD Stage	4

Nutritional recommendations inferred

Number of meals	Kcal BMR	Kcal Intake
5 or 6 meals per day	946	1703

% min carbohydrates	% max carbohydrates	% simple sugars	% proteins	proteins (g)
45.0	50.0	15	12.12	51.6

% min fats	% max fats	% saturated fats	cholesterol (mg)
30	35	10	300

min fibers (g)	max fibers (g)	sodium	calcium (mg)	required BCAA?
21.457	25.0	2.4 grams per day, 6 grams of salt per day	1500	no

Experiment 2: Acceptability Evaluation

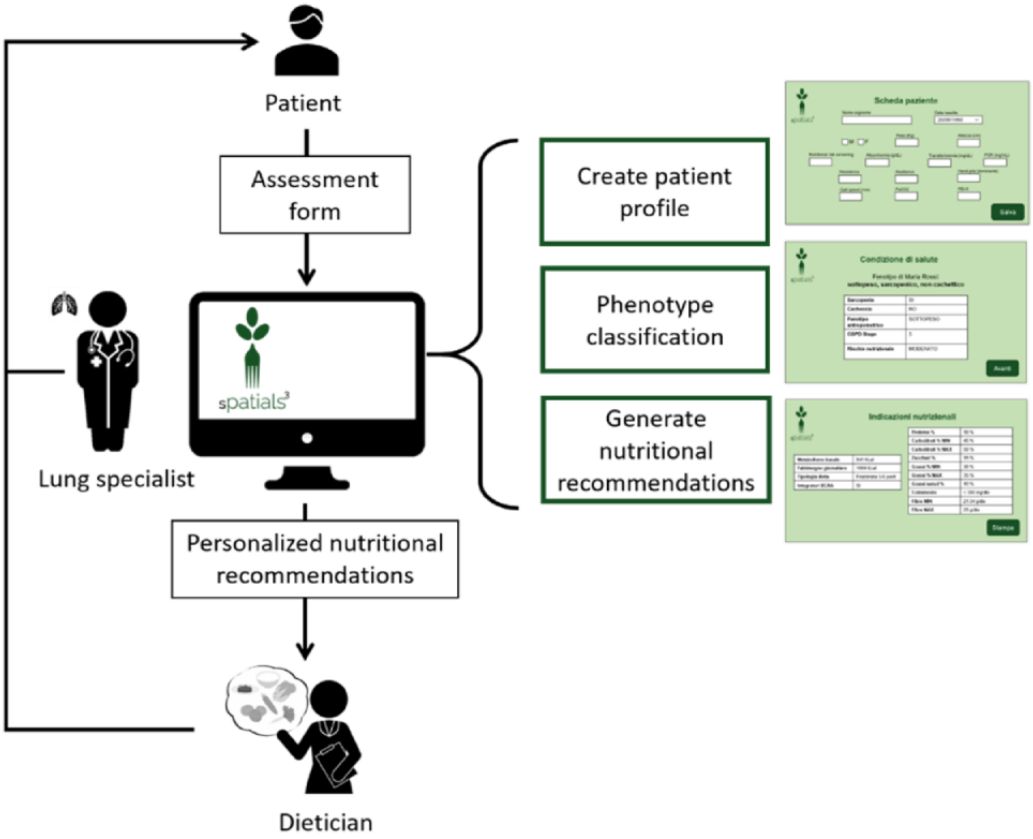
The second evaluation focused on assessing the overall system acceptability, including the DSS and a digital application working as a GUI for inserting patient data and retrieving personalized recommendations. A total of 7 lung specialists agreed to participate in the experiment. The experimenter explained to each participant the expected process of use of the system and the application data flow in a daily clinical routine, as summarized in Figure 5. The lung specialists performed the patient assessment; inserted the clinical data to generate the patient profile on the application GUI, as described in the Application for Clinical Personnel section; and generated the nutritional recommendations**.**

Participants were granted 10 minutes to assess the application, ask the experimenters questions, or ask to be presented with the data flow again.

After this time, the experimenter administered an ad hoc questionnaire based on the subscales of the technology acceptance model by Davis [64] and its subsequent amendments [65] focused on perceived usefulness and intention to use. The participants had to rate on a scale from 1 to 5 the level of agreement with the following statements: (1) "I think that the proposed system is useful for clinicians"; (2) "I think that by using this application could enhance the treatment of individuals with COPD"; and (3) "if I had this application at work, I would use it."

We also asked them to specify at least one reason why the application could be useful or not.

Figure 5. The process of use of the chronic obstructive pulmonary disease decision support system allowing the professionals to generate nutritional recommendations for a specific patient and some screenshots of the main graphical user interface panels (patient profile, patient classification, and nutritional recommendations).



Results

Experiment 1 Results: Validation of the DSS Recommendations

A total of 7 experts (mean age 46.60, SD 13.35 years; n=7, 100% female) participated in the first validation phase. They were all dieticians with an average of 25.57 (SD 11.49) years of professional experience. The level of agreement with each recommendation is reported in Table 6 in terms of mean, SD, and minimum and maximum score.

The comments provided by each participant were analyzed and categorized according to their content. The categorization and frequency of the comments and the patient profiles that originated them are reported in Table 7.

Table 6. Validation—level of agreement scores (mean and SD) for each patient profile (BB, FG, LA, TM, and XY2 are the patient IDs) expressed by each participant; mean and SD for each patient profile.

	Scores (1-5)							Values, mean (SD)
BB	5	4	5	5	3	4	5	4.43 (0.79)
FG	5	1	4	2	3	4	4	3.29 (1.38)
LA	5	4	5	5	4	4	4	4.43 (0.53)
TM	5	4	5	5	3	4	5	4.43 (0.79)
XY2	5	4	5	5	4	4	5	4.57 (0.53)

Table 7. The list of comments provided by clinicians categorized according to their content (category and comment) and linked to the patients that originated them (patient IDs); the frequency of each comment is reported in parentheses.

Category	Comment	Patient ID (frequency)
Quantity	"Too high-calorie uptake"	BB (1), FG (6), TM (1), and ANM (1)
Quantity	"Too high protein uptake"	FG (2) and LA (2)
Quantity	"Consider reducing simple sugars from 15 to 10%"	FG (1), LA (1), TM (2), and ANM (1)
Assessment	"I suggest including the physical activity level during the assessment and adjust correction factors accordingly"	All (2)

Experiment 2 Results: Acceptability Evaluation

A total of 7 experts (mean age 44.71, SD 11.94 years; n=7, 100% female) participated in the second validation phase. Of the 7 experts, 5 (71%) were lung specialists, 1 (14%) was a specialized lung physician, and 1 (14%) was a surgeon of the respiratory system. They had, on average, 14.73 (SD 9.67) years of professional experience, ranging from a minimum of 1 year for the specialized physician to 31 years. The acceptability score for the 3 subscales of perceived usefulness and intention to use was >4 points out of 5, as reported in Table 8.

Our system was considered useful for clinical practice because (1) it promotes the importance and facilitates the inclusion of the nutritional aspect in PR, as stated by 43% (3/7) of the participants; (2) it quickly provides a complete overview of the patient's condition, according to 57% (4/7) of the participants; and (3) it fosters the multidisciplinary collaboration between lung specialists and dieticians. In addition, 86% (6/7) of the participants spontaneously commented on the ease of use of the application GUI that allowed the clinician to insert the patient assessment and obtain the nutritional recommendations.

Table 8. Acceptability—technology acceptance model subscale scores reported by each participant for perceived usefulness (PU1 and PU2) and intention to use (INT); mean and SD are reported for each subscale.

	Scores (1-5)							Values, mean (SD)
INT	4	4	4	5	4	4	4	4.86 (0.38)
PU1	5	5	5	5	4	5	5	4.86 (0.38)
PU2	5	5	5	5	5	5	4	4.14 (0.38)

Discussion

Principal Findings

We performed 2 types of evaluation with 2 different experiments aimed at validating the nutritional recommendations generated by our system by nutrition experts and assessing the system's acceptability by lung specialists. The first validation—performed by 7 nutrition experts—demonstrated that our system is able to provide meaningful and safe recommendations overall in compliance with clinical practice. In 80% (4/5) of the cases, the level of agreement between the human and the "digital" expert was approximately 4.5 points out of 5. In one case (patient FG), the experts did not completely agree with the recommendations, reporting a score of 3.29, which is not considered a disagreement. However, such a score was associated with the case of a critical patient who, in addition to COPD, had second-degree obesity. This is because our system is highly specialized in treating patients with COPD, who need a higher energy intake to cope with impaired respiratory functionality. In such critical cases, the active involvement of the clinician in the process becomes essential. For instance, the clinician may adjust the nutritional recommendations for the patient to lose weight while monitoring them. It is crucial that such a patient does not lose muscular mass instead of fat mass.

The comments were positive overall and could be considered more as suggestions than criticisms. The presence of small disagreements among experts (eg, regarding the calculation of the quantity of macro- and micronutrients) confirms the need for maintaining "the clinician in the loop," as prescribed by the AgiSCOnt methodology adopted for the development of our DSS. In fact, although our system provides a useful and easy way of generating recommendations, each clinical case should be carefully considered, and slight modifications should be made by the clinician themselves in person.

The same considerations apply to the spontaneous comments, summarized in Table 7. The main concerns were about the percentage of simple sugars, which, for 4 patients, could be reduced from 15% to 10%. The guidelines indicate 15% as the maximum value, which should be adjusted by the clinician based on the percentage of other macronutrients. The other comments revealed a slight disagreement on the overall energy and protein intake, which was sometimes considered too high. However, our system is specifically focused on COPD; therefore, a higher intake was justified by the need to compensate for impaired respiratory function. Our group of experts was representative of the Italian clinical scenario, in which the influence of COPD on a patient's nutrition is sometimes neglected. Therefore, such a result strengthens the rationale of our work, which provides a system able to help professionals and clinical care facilities, which often lack specialized services, identify particular needs toward more personalized and effective care.

The second evaluation demonstrated the acceptability of our system by a group of final users (ie, lung specialists involved daily in the assessment and therapy of patients with COPD). The usefulness of our system was confirmed and was especially

related to the possibility of strengthening the consideration of nutritional aspects as part of PR standard practice. This is considered crucial by most specialists and the scientific community; however, due to organizational issues, it is not always considered [4]. Another crucial aspect that emerged was related to the importance of a multidisciplinary approach, and our system could especially help ease the cooperation between lung and nutrition specialists. At the same time, it could help lung professionals in extending their knowledge by considering aspects not strictly related to their expertise. Finally, all participants expressed their willingness to have such a system available in their daily clinical routine. They also considered it easy to use as the GUI was clear and the process was intuitive.

Limitations

This work is not without limitations. First, our system was designed for the Italian context. The DSS is based on the national nutritional guidelines—it was necessary to follow a recognized standard, which may be different from one country to another. Similarly, the dietary recommendations are based on the Italian diet. This was necessary to provide a tool that can be effectively used by our target users (ie, lung specialists treating patients with COPD in the Italian health care system). The DSS's modularity allows it to overcome such a limitation easily—the DSS could be adapted to include nutritional recommendations for other countries. The second limitation is about the participants of our validation experiments. The first experiment was based on the evaluation of 5 patient profiles by a group of nutrition experts. The number of patients examined was identified as the best compromise between a comprehensive representation of the clinical context and organizational aspects. Despite being few, the proposed patient profiles covered most of the potential real clinical cases. Regarding the acceptability evaluation, the main limitation resides in the fact that participants were homogeneous in terms of age (most of them were aged 45-50 years), culture, geographical location, and language (all of them were Italian). Such sociocultural factors are known to impact digital health technology use [66]; however, as previously stated, our work at this stage is focused on the Italian scenario, and therefore, our sample can be considered representative of the final population of target users.

Future Work

As recently noted, most digital health applications remain limited to pilot studies—mainly because they fail in the proposed aims or face significant implementational barriers [67]. From a digital health application perspective, our experiments aimed to verify the stability of the developed solution—in line with the WHO's guidelines [23]. In particular, the validations verified the performance consistency, the proposed solution's overall feasibility, and the digital tool's efficacy. Considering the early stage of the DSS and its application, we need to further investigate the implementation protocols and to acquire long-term proof of the efficacy of the tool among pneumologists and patients with COPD (ie, the acceptability of the tool needs to be verified with a larger sample of end users and in a real clinical setting so that more pneumologists can provide feedback regarding the tool's perceived usefulness, ease of use, and satisfaction; moreover, the effectiveness of the proposed diets should be tested with patients with COPD). To achieve these objectives, more extensive experimentation with larger samples of participants (clinical nutritionists, dieticians, and pneumologists) is necessary. Moreover, the involvement of clinical personnel can support the identification of implementation protocols suitable for the adoption of the digital tool in clinical practice. In this way, the application's level of maturity could move from early to mild (according to the WHO [23]), where its effectiveness can be tested in a nonresearch (uncontrolled) setting.

To support the prompt identification of barriers and implementational challenges, reporting the development of the COPD DSS within a framework for the definition of digital health application implementation can be useful. In particular, the Guidelines and Checklist for the Reporting on Digital Health Implementations [67], by providing a list of 20 items to be monitored, can foster the identification of issues in the Technical design phase in the Interoperability and Data management areas.

In this regard, the availability of data and the implementation of the application within the health system are another relevant node to be investigated. Although the current version of the digital application is still in its prototypical phase, scaling up the application to the regional or national level (ie, *coverage* in the Guidelines and Checklist for the Reporting on Digital Health Implementations) is essential to ensure its use in clinical practice. Therefore, toward this aim, strategies for collecting the outputs and making them available in patients' data (or electronic health records) need to be investigated. In this regard, scientific literature offers some interesting approaches grounded in the Italian health care system that could be considered to make the COPD DSS interoperable with existing tools [68-70]. As the COPD DSS leverages ontologies to represent its data, this technology can be used to achieve semantic interoperability of the information [68], moving a step toward the longitudinal collection of health data about patients [70] while ensuring data protection according to the national and European laws [69].

Finally, from an ontological perspective, the domain ontology regarding COPD presented in this work could benefit from mapping with existing (and larger) biomedical standard ontologies to increase its shareability (eg, the WHO's International Classification of Diseases and International Classification of Functioning, Disability and Health, as well as upper biomedical ontologies)—a best practice of ontology engineering [71]. Moreover, considering that the proposed system can be adapted to any other national clinical context by modifying the domain ontologies, a possible future research direction consists of involving international clinicians to increase the knowledge formalized in the ontologies so that it is possible for the DSS to cover the specific nutritional indications of different countries.

Conclusions

The role of nutrition in the management of patients with COPD is often underestimated, although scientific evidence points toward the important role that diet plays in PR. The nutritional status of patients with COPD is essential to prevent exacerbations and avoid comorbidities, but attention to the patient's body composition and nutritional status is often

secondary in clinical practice. This may be partially because lung specialists may lack specialized training in clinical nutrition, although they recognize the relevance of dietary recommendations in PR.

An ontology-based DSS was developed to support pneumologists in considering nutritional aspects. The DSS formalizes expert knowledge in computable models able to infer patient-tailored nutritional recommendations, leveraging a set of information to capture the nutritional and physical status of the patient; therefore, by applying rules, it can support the classification of patients with COPD and provide tailored recommendations indicating the percentages and amounts of micro- and macronutrients that should make up their diet. The domain ontologies act as the backbone of a clinician-dedicated application.

The application was validated to assess the clinical compliance of the DSS's recommendations and the acceptability of such an application in clinical practice by lung specialists. For both validations, the proposed system performed more than adequately—in particular, pneumologists underlined the role that such an application may play in achieving a multidisciplinary approach in PR. This paper concludes by investigating future research directions to implement the COPD DSS further into a fully-fledged digital health application.

Acknowledgments

The sPATIALS3 (Miglioramento delle Produzioni Agroalimentari e Tecnologie Innovative per un'Alimentazione più Sana, Sicura e Sostenibile) project is financed by the European Regional Development Fund under the ROP (Regional Operational Programme) of the Lombardy Region European Regional Development Fund 2014 to 2020 (axis I, "Strengthen technological research, development and innovation"; action 1.b.1.3, "Support for cooperative R&D activities to develop new sustainable technologies, products and services"; Call Hub).

Data Availability

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

Conflicts of Interest

None declared.

Multimedia Appendix 1
A table representing the set of rules for providing phenotype-tailored recommendations.
[DOCX File , 46 KB - medinform_v12i1e50980_app1.docx]

Multimedia Appendix 2
The chronic obstructive pulmonary disease decision support system domain ontology.
[ZIP File (Zip Archive), 24 KB - medinform_v12i1e50980_app2.zip]

Multimedia Appendix 3
A table reporting the nutritional recommendations for each patient used to test the validity of the ontology rule set.
[DOCX File , 31 KB - medinform_v12i1e50980_app3.docx]

References

1. Benjafield A, Tellez D, Barrett M, Gondalia R, Nunez C, Wedzicha J, et al. An estimate of the European prevalence of COPD in 2050. Eur Respir J 2021;58(suppl 65):OA2866 [FREE Full text] [doi: 10.1183/13993003.congress-2021.oa2866]
2. MacNee W. Pathology, pathogenesis, and pathophysiology. BMJ 2006 May 18;332(7551):1202-1204. [doi: 10.1136/bmj.332.7551.1202]
3. Spruit MA, Singh SJ, Garvey C, Zu Wallack R, Nici L, Rochester C, ATS/ERS Task Force on Pulmonary Rehabilitation. An official American thoracic society/European respiratory society statement: key concepts and advances in pulmonary rehabilitation. Am J Respir Crit Care Med 2013 Oct 15;188(8):e13-e64. [doi: 10.1164/rccm.201309-1634ST] [Medline: 24127811]
4. Holland AE, Cox NS, Houchen-Wolloff L, Rochester CL, Garvey C, ZuWallack R, et al. Defining modern pulmonary rehabilitation: an official American thoracic society workshop report. Ann Am Thorac Soc 2021 May;18(5):e12-e29 [FREE Full text] [doi: 10.1513/AnnalsATS.202102-146ST] [Medline: 33929307]
5. Scoditti E, Massaro M, Garbarino S, Toraldo DM. Role of diet in chronic obstructive pulmonary disease prevention and treatment. Nutrients 2019 Jun 16;11(6):1357 [FREE Full text] [doi: 10.3390/nu11061357] [Medline: 31208151]
6. Collins PF, Stratton RJ, Elia M. Nutritional support in chronic obstructive pulmonary disease: a systematic review and meta-analysis. Am J Clin Nutr 2012 Jun;95(6):1385-1395 [FREE Full text] [doi: 10.3945/ajcn.111.023499] [Medline: 22513295]

7. Ferreira I, Brooks D, White J, Goldstein R. Nutritional supplementation for stable chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2012 Dec 12;12:CD000998. [doi: 10.1002/14651858.CD000998.pub3] [Medline: 23235577]
8. Collins PF, Yang IA, Chang YC, Vaughan A. Nutritional support in chronic obstructive pulmonary disease (COPD): an evidence update. J Thorac Dis 2019 Oct;11(Suppl 17):S2230-S2237 [FREE Full text] [doi: 10.21037/jtd.2019.10.41] [Medline: 31737350]
9. Liu H, Tan X, Liu Z, Ma X, Zheng Y, Zhu B, et al. Association between diet-related inflammation and COPD: findings from NHANES III. Front Nutr 2021 Oct 18;8:732099 [FREE Full text] [doi: 10.3389/fnut.2021.732099] [Medline: 34733875]
10. van Iersel LE, Beijers RJ, Gosker HR, Schols AM. Nutrition as a modifiable factor in the onset and progression of pulmonary function impairment in COPD: a systematic review. Nutr Rev 2022 May 09;80(6):1434-1444 [FREE Full text] [doi: 10.1093/nutrit/nuab077] [Medline: 34537848]
11. Nguyen HT, Collins PF, Pavey TG, Nguyen NV, Pham TD, Gallegos DL. Nutritional status, dietary intake, and health-related quality of life in outpatients with COPD. Int J Chron Obstruct Pulmon Dis 2019;14:215-226 [FREE Full text] [doi: 10.2147/COPD.S181322] [Medline: 30666102]
12. van Bakel SI, Gosker HR, Langen RC, Schols AM. Towards personalized management of sarcopenia in COPD. Int J Chron Obstruct Pulmon Dis 2021;16:25-40 [FREE Full text] [doi: 10.2147/COPD.S280540] [Medline: 33442246]
13. Schols AM, Ferreira IM, Franssen FM, Gosker HR, Janssens W, Muscaritoli M, et al. Nutritional assessment and therapy in COPD: a European respiratory society statement. Eur Respir J 2014 Dec 18;44(6):1504-1520 [FREE Full text] [doi: 10.1183/09031936.00070914] [Medline: 25234804]
14. Machado FV, Schneider LP, Fonseca J, Belo LF, Bonomo C, Morita AA, et al. Clinical impact of body composition phenotypes in patients with COPD: a retrospective analysis. Eur J Clin Nutr 2019 Nov 14;73(11):1512-1519. [doi: 10.1038/s41430-019-0390-4] [Medline: 30643222]
15. Hanson C, Rutten EP, Wouters EF, Rennard S. Influence of diet and obesity on COPD development and outcomes. Int J Chron Obstruct Pulmon Dis 2014;9:723-733 [FREE Full text] [doi: 10.2147/COPD.S50111] [Medline: 25125974]
16. Mete B, Pehlivan E, Gülbaş G, Günen H. Prevalence of malnutrition in COPD and its relationship with the parameters related to disease severity. Int J Chron Obstruct Pulmon Dis 2018;13:3307-3312 [FREE Full text] [doi: 10.2147/COPD.S179609] [Medline: 30349235]
17. Sepúlveda-Loyola W, Osadnik C, Phu S, Morita AA, Duque G, Probst VS. Diagnosis, prevalence, and clinical impact of sarcopenia in COPD: a systematic review and meta-analysis. J Cachexia Sarcopenia Muscle 2020 Oct 30;11(5):1164-1176 [FREE Full text] [doi: 10.1002/jcsm.12600] [Medline: 32862514]
18. Hoong JM, Ferguson M, Hukins C, Collins PF. Economic and operational burden associated with malnutrition in chronic obstructive pulmonary disease. Clin Nutr 2017 Aug;36(4):1105-1109. [doi: 10.1016/j.clnu.2016.07.008] [Medline: 27496063]
19. Iheanacho I, Zhang S, King D, Rizzo M, Ismaila AS. Economic burden of chronic obstructive pulmonary disease (COPD): a systematic literature review. Int J Chron Obstruct Pulmon Dis 2020;15:439-460 [FREE Full text] [doi: 10.2147/COPD.S234942] [Medline: 32161455]
20. Beijers RJ, Steiner MC, Schols AM. The role of diet and nutrition in the management of COPD. Eur Respir Rev 2023 Jun 30;32(168):230003 [FREE Full text] [doi: 10.1183/16000617.0003-2023] [Medline: 37286221]
21. Raad S, Smith C, Allen K. Nutrition status and chronic obstructive pulmonary disease: can we move beyond the body mass index? Nutr Clin Pract 2019 Jun;34(3):330-339. [doi: 10.1002/ncp.10306] [Medline: 30989731]
22. Watson A, Wilkinson TM. Digital healthcare in COPD management: a narrative review on the advantages, pitfalls, and need for further research. Ther Adv Respir Dis 2022 Mar 02;16:17534666221075493 [FREE Full text] [doi: 10.1177/17534666221075493] [Medline: 35234090]
23. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. 2016. URL: https://www.who.int/publications/i/item/9789241511766 [accessed 2024-04-29]
24. Spoladore D, Colombo V, Arlati S, Mahroo A, Trombetta A, Sacco M. An ontology-based framework for a telehealthcare system to foster healthy nutrition and active lifestyle in older adults. Electronics 2021 Sep 01;10(17):2129. [doi: 10.3390/electronics10172129]
25. Greenberg J, Deshmukh R, Huang L, Mostafa J, La Vange L, Carretta E, et al. The COPD ontology and toward empowering clinical scientists as ontology engineers. J Libr Metadata 2010 Aug 31;10(2-3):173-187. [doi: 10.1080/19386389.2010.520604]
26. Cano I, Tényi Á, Schueller C, Wolff M, Huertas Migueláñez MM, Gomez-Cabrero D, et al. The COPD knowledge base: enabling data analysis and computational simulation in translational COPD research. J Transl Med 2014;12(Suppl 2):S6. [doi: 10.1186/1479-5876-12-s2-s6]
27. Rayner L, Sherlock J, Creagh-Brown B, Williams J, deLusignan S. The prevalence of COPD in England: an ontological approach to case detection in primary care. Respir Med 2017 Nov;132:217-225 [FREE Full text] [doi: 10.1016/j.rmed.2017.10.024] [Medline: 29229101]
28. Rosso R, Munaro G, Salvetti O, Colantonio S, Ciancitto F. CHRONIOUS: an open, ubiquitous and adaptive chronic disease management platform for chronic obstructive pulmonary disease (COPD), chronic kidney disease (CKD) and renal insufficiency. Annu Int Conf IEEE Eng Med Biol Soc 2010;2010:6850-6853. [doi: 10.1109/IEMBS.2010.5626451] [Medline: 21096301]

29. Lasierra N, Alesanco A, Guillén S, García J. A three stage ontology-driven solution to provide personalized care to chronic patients at home. J Biomed Inform 2013 Jun;46(3):516-529 [FREE Full text] [doi: 10.1016/j.jbi.2013.03.006] [Medline: 23567539]
30. Ajami H, McHeick H. Ontology-based model to support ubiquitous healthcare systems for COPD patients. Electronics 2018 Dec 02;7(12):371. [doi: 10.3390/electronics7120371]
31. McCabe C, McCann M, Brady AM. Computer and mobile technology interventions for self-management in chronic obstructive pulmonary disease. Cochrane Database Syst Rev 2017 May 23;5(5):CD011425 [FREE Full text] [doi: 10.1002/14651858.CD011425.pub2] [Medline: 28535331]
32. Colombo V, Mondellini M, Gandolfo A, Fumagalli A, Sacco M. A mobile diary app to support rehabilitation at home for elderly with COPD: a preliminary feasibility study. In: Proceedings of the 17th International Conference on Computers Helping People with Special Needs. 2020 Presented at: ICCHP '20; September 9-11, 2020; Lecco, Italy p. 224-232 URL: https://dl.acm.org/doi/abs/10.1007/978-3-030-58805-2_27 [doi: 10.1007/978-3-030-58805-2_27]
33. myCOPD - empowering patients to manage their COPD for a lifetime. my mhealth Limited. URL: https://mymhealth.com/mycopd [accessed 2024-04-29]
34. The problem with AI. Earley Information Science. 2017. URL: https://www.earley.com/insights/problem-with-ai [accessed 2024-04-29]
35. Spoladore D, Sacco M, Trombetta A. A review of domain ontologies for disability representation. Expert Syst Appl 2023 Oct;228:120467. [doi: 10.1016/j.eswa.2023.120467]
36. Kotis KI, Vouros GA, Spiliotopoulos D. Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations. Knowl Eng Rev 2020 Jan 31;35:e4 [FREE Full text] [doi: 10.1017/s0269888920000065]
37. Spoladore D, Pessot E, Trombetta A. A novel agile ontology engineering methodology for supporting organizations in collaborative ontology development. Comput Ind 2023 Oct;151:103979. [doi: 10.1016/j.compind.2023.103979]
38. Owen T. Building expert systems, edited by Frederick Hayes-Roth, Donald A. Waterman and Douglas B. Lenat Addison-Wesley Publishing Company, Massachusetts, USA, 1983 (£32.95). Robotica 2009 Mar 09;6(2):165. [doi: 10.1017/s0263574700004069]
39. Grüninger M, Fox MS. The role of competency questions in enterprise engineering. In: Rolstadås A, editor. Benchmarking — Theory and Practice. Cham, Switzerland: Springer; 1995:22-31.
40. Gangemi A. Ontology design patterns for semantic web content. In: Proceedings of the 4th International Semantic Web Conference on Semantic Web. 2005 Presented at: ISWC '05; November 6-10, 2005; Galway, Ireland p. 262-276 URL: https://link.springer.com/chapter/10.1007/11574620_21 [doi: 10.1007/11574620_21]
41. Evans WJ, Morley JE, Argilés J, Bales C, Baracos V, Guttridge D, et al. Cachexia: a new definition. Clin Nutr 2008 Dec;27(6):793-799. [doi: 10.1016/j.clnu.2008.06.013] [Medline: 18718696]
42. Agustí A, Celli BR, Criner GJ, Halpin D, Anzueto A, Barnes P, et al. Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary. Eur Respir J 2023 Apr;61(4):2300239 [FREE Full text] [doi: 10.1183/13993003.00239-2023] [Medline: 36858443]
43. Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2013 Feb 15;187(4):347-365. [doi: 10.1164/rccm.201204-0596pp]
44. Prendergast JM, Coe RM, Chavez MN, Romeis JC, Miller DK, Wolinsky FD. Clinical validation of a nutritional risk index. J Community Health 1989;14(3):125-135. [doi: 10.1007/BF01324362] [Medline: 2600200]
45. Sergi G, De Rui M, Veronese N, Bolzetta F, Berton L, Carraro S, et al. Assessing appendicular skeletal muscle mass with bioelectrical impedance analysis in free-living Caucasian older adults. Clin Nutr 2015 Aug;34(4):667-673. [doi: 10.1016/j.clnu.2014.07.010] [Medline: 25103151]
46. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, Writing Group for the European Working Group on Sarcopenia in Older People 2 (EWGSOP2), the Extended Group for EWGSOP2. Sarcopenia: revised European consensus on definition and diagnosis. Age Ageing 2019 Jan 01;48(1):16-31 [FREE Full text] [doi: 10.1093/ageing/afy169] [Medline: 30312372]
47. LARN - Livelli di Assunzione di Riferimento di Nutrienti ed energia per la popolazione italiana. Società Italiana di Nutrizione Umana (SINU). 2021. URL: https://sinu.it/tabelle-larn-2014/ [accessed 2024-06-03]
48. Frankenfield D, Roth-Yousey L, Compher C. Comparison of predictive equations for resting metabolic rate in healthy nonobese and obese adults: a systematic review. J Am Diet Assoc 2005 May;105(5):775-789. [doi: 10.1016/j.jada.2005.02.005] [Medline: 15883556]
49. Bendavid I, Lobo DN, Barazzoni R, Cederholm T, Coëffier M, de van der Schueren M, et al. The centenary of the Harris-Benedict equations: how to assess energy requirements best? Recommendations from the ESPEN expert group. Clin Nutr 2021 Mar;40(3):690-701. [doi: 10.1016/j.clnu.2020.11.012] [Medline: 33279311]
50. Pitta F, Troosters T, Spruit MA, Probst VS, Decramer M, Gosselink R. Characteristics of physical activities in daily life in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2005 May 01;171(9):972-977. [doi: 10.1164/rccm.200407-855OC] [Medline: 15665324]

51. Matheson EM, Nelson JL, Baggs GE, Luo M, Deutz NE. Specialized oral nutritional supplement (ONS) improves handgrip strength in hospitalized, malnourished older patients with cardiovascular and pulmonary disease: a randomized clinical trial. Clin Nutr 2021 Mar;40(3):844-849. [doi: 10.1016/j.clnu.2020.08.035] [Medline: 32943241]
52. Bai GH, Tsai MC, Tsai HW, Chang CC, Hou WH. Effects of branched-chain amino acid-rich supplementation on EWGSOP2 criteria for sarcopenia in older adults: a systematic review and meta-analysis. Eur J Nutr 2022 Mar 27;61(2):637-651. [doi: 10.1007/s00394-021-02710-0] [Medline: 34705076]
53. Guerra BA, Pereira TG, Eckert IC, Bernardes S, Silva FM. Markers of respiratory function response to high-carbohydrate and high-fat intake in patients with lung diseases: a systematic review with meta-analysis of randomized clinical trials. JPEN J Parenter Enteral Nutr 2022 Sep 31;46(7):1522-1534. [doi: 10.1002/jpen.2385] [Medline: 35437762]
54. WHO global report on sodium intake reduction. World Health Organization.: World Health Organization; 2023. URL: https://www.who.int/publications/i/item/9789240069985 [accessed 2024-04-29]
55. Chen Y, Ramsook AH, Coxson HO, Bon J, Reid WD. Prevalence and risk factors for osteoporosis in individuals with COPD: a systematic review and meta-analysis. Chest 2019 Dec;156(6):1092-1110. [doi: 10.1016/j.chest.2019.06.036] [Medline: 31352034]
56. 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]
57. Pan JZ. Resource description framework. In: Staab S, Studer R, editors. Handbook on Ontologies. Berlin, Germany: Springer; 2009:71-90.
58. Antoniou G, van Harmelen F. Web ontology language: OWL. In: Antoniou G, van Harmelen F, editors. Handbook on Ontologies. Cham, Switzerland: Springer; 2009:91-110.
59. Alamri A, Bertok P. Distributed store for ontology data management. In: Lee R, editor. Computer and Information Science. Cham, Switzerland: Springer; 2012:15-35.
60. Spoladore D, Mahroo A, Trombetta A, Sacco M. DOMUS: a domestic ontology managed ubiquitous system. J Ambient Intell Human Comput 2021 Mar 31;13(6):3037-3052. [doi: 10.1007/S12652-021-03138-4]
61. O'Connor M, Tu S, Nyulas C, Das A, Musen M. Querying the semantic web with SWRL. In: Proceedings of the 2007 International Symposium on Advances in Rule Interchange and Applications.: Springer; 2007 Presented at: RuleML '07; October 25-26, 2007; Orlando, FL p. 155-159 URL: https://link.springer.com/chapter/10.1007/978-3-540-75975-1_13 [doi: 10.1007/978-3-540-75975-1_13]
62. Hommeaux EP, Seaborne A. SPARQL query language for RDF. W3C Recommendation. 2008. URL: https://www.w3.org/TR/rdf-sparql-query/ [accessed 2024-04-29]
63. Spoladore D, Mahroo A, Sacco M. Leveraging ontology to enable indoor comfort customization in the smart home. In: Proceedings of the 13th International Conference on Flexible Query Answering Systems. 2019 Presented at: FQAS '19; July 2-5, 2019; Amantea, Italy p. 63-74 URL: https://link.springer.com/chapter/10.1007/978-3-030-27629-4_9 [doi: 10.1007/978-3-030-27629-4_9]
64. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340 [FREE Full text] [doi: 10.2307/249008]
65. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
66. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
67. 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]
68. Ciampi M, Esposito A, Guarasci R, De Pietro G. Towards interoperability of EHR systems: the case of Italy. In: Proceedings of the 2nd International Conference on Information and Communication Technologies for Ageing Well and e-Health. 2016 Presented at: ICT4AGEINGWELL '16; April 21-22, 2016; Rome, Italy p. 138 URL: https://www.scitepress.org/Link.aspx?doi=10.5220/0005916401330138 [doi: 10.5220/0005916401330138]
69. Bologna S, Bellavista A, Corso PP, Zangara G. Electronic health record in Italy and personal data protection. Eur J Health Law 2016 Jun 14;23(3):265-277. [doi: 10.1163/15718093-12341403] [Medline: 27491249]
70. Ciampi M, Sicuranza M, Esposito A, Guarasci R, De Pietro G. A technological framework for EHR interoperability: experiences from Italy. In: Proceedings of the 2nd International Conference on Information and Communication Technologies for Ageing Well and e-Health. 2016 Presented at: ICT4AWE '16; April 21-22, 2016; Rome, Italy p. 80-99 URL: https://link.springer.com/chapter/10.1007/978-3-319-62704-5_6 [doi: 10.1007/978-3-319-62704-5_6]
71. Spoladore D, Pessot E. Collaborative ontology engineering methodologies for the development of decision support systems: case studies in the healthcare domain. Electronics 2021 Apr 29;10(9):1060. [doi: 10.3390/electronics10091060]

Abbreviations

AgiSCOnt: Agile, Simplified, and Collaborative Ontology Engineering Methodology

BCAA: branched-chain amino acid
BMR: basal metabolic rate
COPD: chronic obstructive pulmonary disease
CQ: competency question
DSS: decision support system
FEV1: forced expiratory volume in the first second
FFM: fat-free mass
GUI: graphical user interface
LARN: Livelli di Assunzione di Riferimento di Nutrienti ed energia
NRI: nutritional risk index
PR: pulmonary rehabilitation
SWRL: semantic web rule language
WHO: World Health Organization

Edited by C Lovis; submitted 18.07.23; peer-reviewed by A AL-Asadi, T Salzmann; comments to author 19.01.24; revised version received 01.02.24; accepted 23.04.24; published 26.06.24.

Please cite as:
Spoladore D, Colombo V, Fumagalli A, Tosi M, Lorenzini EC, Sacco M
An Ontology-Based Decision Support System for Tailored Clinical Nutrition Recommendations for Patients With Chronic Obstructive Pulmonary Disease: Development and Acceptability Study
JMIR Med Inform 2024;12:e50980
URL: https://medinform.jmir.org/2024/1/e50980
doi:10.2196/50980
PMID:38922666

Use of Video in Telephone Triage in Out-of-Hours Primary Care: Register-Based Study

Mette Amalie Nebsbjerg[1], MD; Claus Høstrup Vestergaard[1], MSc; Katrine Bjørnshave Bomholt[1], MD; Morten Bondo Christensen[1,2], MD, PhD; Linda Huibers[1], MD, PhD

1

2

Corresponding Author:
Mette Amalie Nebsbjerg, MD

Abstract

Background: Out-of-hours primary care (OOH-PC) is challenging due to high workloads, workforce shortages, and long waiting and transportation times for patients. Use of video enables triage professionals to visually assess patients, potentially ending more contacts in a telephone triage contact instead of referring patients to more resource-demanding clinic consultations or home visits. Thus, video use may help reduce use of health care resources in OOH-PC.

Objective: This study aimed to investigate video use in telephone triage contacts to OOH-PC in Denmark by studying rate of use and potential associations between video use and patient- and contact-related characteristics and between video use and triage outcomes and follow-up contacts. We hypothesized that video use could serve to reduce use of health care resources in OOH-PC.

Methods: This register-based study included all telephone triage contacts to OOH-PC in 4 of the 5 Danish regions from March 15, 2020, to December 1, 2021. We linked data from the OOH-PC electronic registration systems to national registers and identified telephone triage contacts with video use (video contact) and without video use (telephone contact). Calculating crude incidence rate ratios and adjusted incidence rate ratios (aIRRs), we investigated the association between patient- and contact-related characteristics and video contacts and measured the frequency of different triage outcomes and follow-up contacts after video contact compared to telephone contact.

Results: Of 2,900,566 identified telephone triage contacts to OOH-PC, 9.5% (n=275,203) were conducted as video contacts. The frequency of video contact was unevenly distributed across patient- and contact-related characteristics; it was used more often for employed young patients without comorbidities who contacted OOH-PC more than 4 hours before the opening hours of daytime general practice. Compared to telephone contacts, notably more video contacts ended with advice and self-care (aIRR 1.21, 95% CI 1.21-1.21) and no follow-up contact (aIRR 1.08, 95% CI 1.08-1.09).

Conclusions: This study supports our hypothesis that video contacts could reduce use of health care resources in OOH-PC. Video use lowered the frequency of referrals to a clinic consultation or a home visit and also lowered the frequency of follow-up contacts. However, the results could be biased due to confounding by indication, reflecting that triage GPs use video for a specific set of reasons for encounters.

(JMIR Med Inform 2024;12:e47039) doi:10.2196/47039

KEYWORDS

primary health care; after-hours care; referral and consultation; general practitioner; GP; triage; remote consultation; telemedicine

Introduction

General practice serves as a gatekeeper to secondary care in many countries [1]. However, the services in out-of-hours primary care (OOH-PC) are challenging due to high workloads, workforce shortages, and long waiting and transportation times for patients. This development has received much political attention and has caused public debate and reorganization [2,3].

Existing health care systems are currently undergoing a digital transformation, which was pushed by the COVID-19 pandemic [4-9]. As a central part of this digitization, video consultations have been implemented broadly in general practice [5,8-12]. Many countries have introduced video as part of telephone triage in OOH-PC [12-14]. Video use enables triage professionals to visually assess patients, which may imply that more contacts can be ended in a telephone triage contact instead of referring patients to clinic consultations or home visits, which demand more resources. Thereby, video use might reduce use of health care resources related to clinic consultations and home visits.

Research has shown that patients welcome the use of video in general practice in the daytime and also after hours [4,14-16]. However, in daytime general practice, general practitioners (GPs) experience both benefits of (eg, care delivery) and barriers to (eg, technical difficulties, varying suitability for different

health problems and patient groups) video use [6,10,16-19]. Two qualitative studies indicated that video use in OOH-PC is beneficial to both triage professionals (eg, it improved patient assessment and reassurance) [13,14] and patients (eg, it led to better reassurance and higher satisfaction) [14]. Two register-based studies found that video use in OOH-PC increased during the COVID-19 pandemic [12,20]. However, little is still known about video use and its effects. This study aimed to investigate video use in telephone triage contacts to OOH-PC in Denmark by studying rate of use and potential associations between video use and patient- and contact-related characteristics and between video use and triage outcomes and follow-up contacts.

Methods

Design and Population

We conducted a register-based study of video use in telephone triage contacts to OOH-PC in 4 of the 5 Danish regions (North Denmark Region, Central Denmark Region, Region of Southern Denmark, and Region Zealand). As the Capital Region of Denmark runs a different OOH-PC system than the other 4 regions, this region was not included in this study. We included all telephone contacts from March 15, 2020, to December 1, 2021, and followed each patient for 7 days to record the outcomes. In Region Zealand, telephone contacts were included from March 1, 2021, because this region started using video from this date.

Setting

Denmark has free public health care for its residents. The health care system is centrally regulated, but most services are provided by the local governments of the 5 regions. Outside office hours, Danish GPs and GP trainees cover shifts in the regional OOH-PC service, which is open on weekdays from 4 PM to 8 AM and 24 hours during weekends and holidays. GPs and GP trainees in their last year of specialist training (hereinafter referred to jointly as triage GPs) perform telephone triage and determine the triage outcome: telephone triage with video use (video contacts) or telephone triage without video use (telephone contacts), clinic consultation, home visit, or hospital admission. The triage GPs assesses whether the problem is suitable for a video contact. If so and if the patient approves, a video link is sent to the patient via text message. When the link is activated, the triage GP can see the patient, but the patient cannot see the triage GP. Triage GPs are paid a fee for service using remuneration codes.

Outcome Measures

The following outcome measures were defined: the proportion of video contacts (number of video contacts per 100 telephone contacts); the association between video contact and patient- and contact-related characteristics (sex and age of the patient, cohabitation status, comorbidity, educational level, ethnicity, income, urbanization, employment status, region, and time of contact); the frequency of triage outcomes (advice and self-care, referral to clinic consultation, home visit, or hospital admission) and their association with video contact; and the frequency of follow-up contacts in daytime general practice or OOH-PC within 7 days or a hospital admission within 1 day and their association with video contact.

Data Collection

We used data from the OOH-PC electronic registration system, which provided information on date, time, region, type of contact (telephone contact or video contact), and triage outcome (advice and self-care, referral to clinic consultation, home visit, or hospital admission). We constructed a "time of contact" variable, which was defined by its relation to the next opening time of daytime general practice and dichotomized into >4 hours or ≤4 hours, as the option to refer a patient to their regular GP may influence the triage decision.

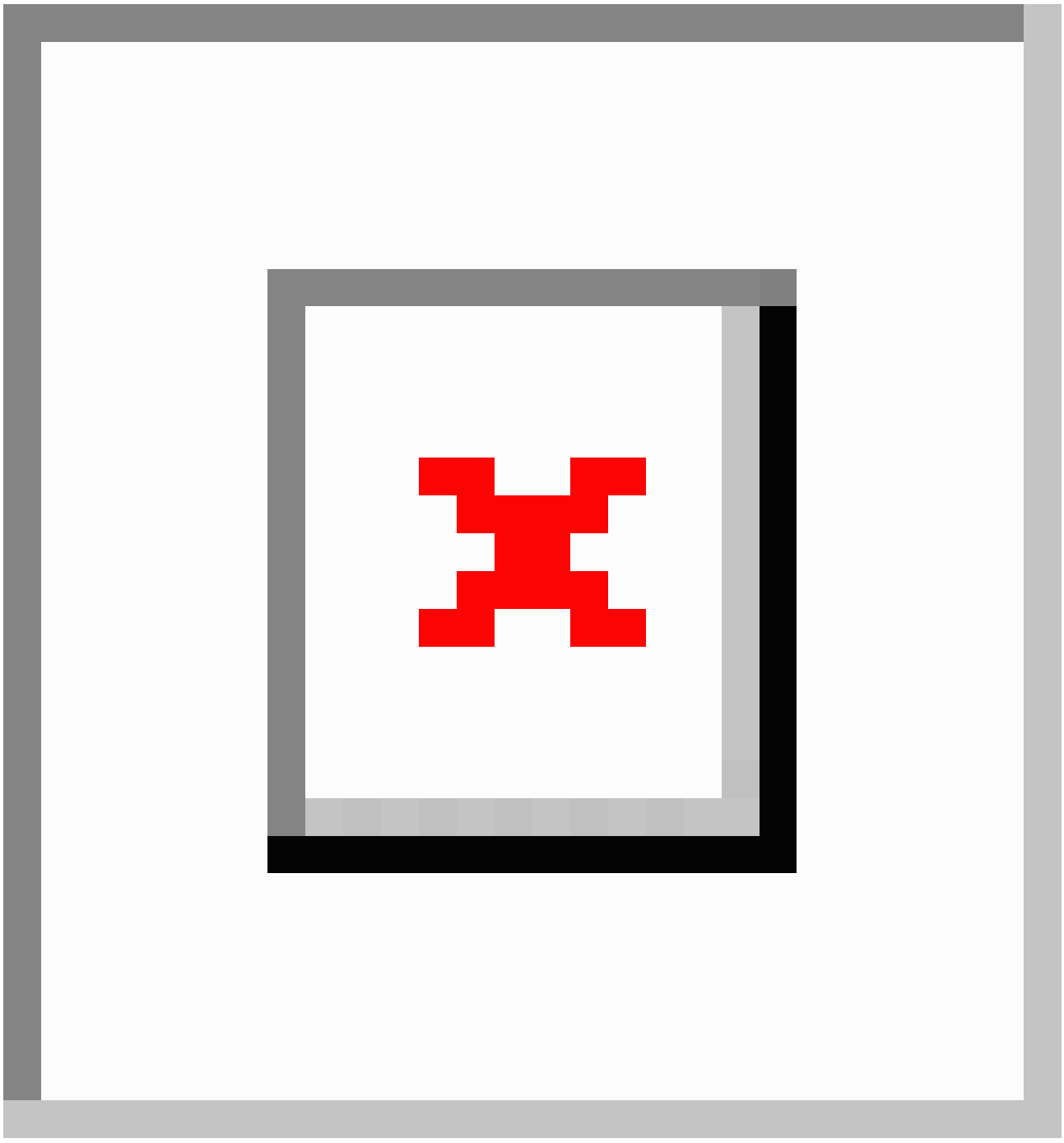
To investigate follow-up contacts, we linked data from the OOH-PC registration system to 2 Danish national registers using each patient's unique personal identification number [21]. The Danish National Health Service Register [22] provided information on date and type of contact to daytime general practice (telephone contacts, video contacts, clinic consultations, or home visits). The Danish National Patient Registry [23] provided information on date of contact to the hospital (emergency department visits and unscheduled hospital admissions) and comorbidity. Comorbidity was defined as the number of diagnoses from the Charlson Comorbidity Index that were recorded as diagnosis codes in hospital charts. Data on socioeconomic characteristics of the patients (sex, age, cohabitation status, educational level, ethnicity, income, urbanization, and employment status) were obtained from Statistics Denmark [24]. All covariates (except for age, sex, and comorbidity) were reported at the household level. For example, household educational level was determined by the member with the longest education. Hence, it was possible to avoid excluding contacts involving children because of missing values. We included only persons with registered socioeconomic characteristics.

Data Analyses

People with more than 25 contacts to OOH-PC during the study period (comprising 98,126/2,900,566 contacts, 3.4%) were excluded from the data analyses since they were considered outliers. Likewise, people aged >104 years (162/2,900,566 contacts, 0%) and patients with missing covariates (18,740/2,900,566 contacts, 0.7%) were excluded.

Descriptive analyses were used to describe the study population. To ensure convergence of the regressions, we used Poisson regression models to measure the association between patient- and contact-related characteristics and video contacts, and we calculated incidence rate ratios (IRRs) and 95% CIs [25]. Results are presented as a forest plot (Figure 1). Using a Poisson regression model, we also calculated crude and adjusted IRRs (aIRRs) of triage outcomes and follow-up contacts after a video contact compared to after a telephone contact. IRRs were adjusted for patient- and contact-related characteristics. Stata (version 17; StataCorp) was used to analyze all data. Reporting of results was conducted in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement.

Figure 1. Forest plot presenting the association between patient- and contact-related characteristics and the likelihood of having a video contact (incidence rate ratios [IRRs] with 95% CIs). An IRR >1 indicates a higher use of video contacts compared to the reference group, marked in the right column of the figure. Conversely, an IRR <1 indicates a lower use of video contacts.



Ethical Considerations

The Committee on Health Research Ethics in the Central Denmark Region approved the data collection from the electronic patient records in the OOH-PC registration system (1-45-70-22-22) without informed consent from participants or any provision for them to opt out. The study was listed in the record of processing activities at the Research Unit for General Practice in Aarhus in accordance with the provisions of the General Data Protection Regulation (GDPR). All data on participants were deidentified. Finally, conduct of the study was endorsed by the regional association of GPs.

Results

Study Population

During the study period, 2,900,566 telephone triage contacts to OOH-PC were identified (Table 1). Patient- and contact-related characteristics varied between telephone and video contacts; the largest variation was seen for patient age, comorbidity, employment status, region, and time of contact.

Table . Distribution of patient- and contact-related characteristics (N=2,900,566).

Characteristics		Telephone contacts (n=2,625,363, 90.5%), n (%)	Video contacts (n=275,203, 9.5%), n (%)	Total (n=2,900,566, 100%), n (%)
Sex				
	Female	1,427,918 (54.4)	140,684 (51.1)	1,568,602 (54.1)
	Male	1,197,445 (45.6)	134,519 (48.9)	1,331,964 (45.9)
Age (years)				
	0-4	311,749 (11.9)	83,672 (30.4)	395,421 (13.6)
	5-10	138,745 (5.3)	25,196 (9.2)	163,941 (5.7)
	11-20	310,179 (11.8)	40,830 (14.8)	351,009 (12.1)
	21-40	688,182 (26.2)	64,780 (23.5)	752,962 (26)
	41-60	508,558 (19.4)	40,955 (14.9)	549,513 (18.9)
	61-80	433,400 (16.5)	16,042 (5.8)	449,442 (15.5)
	≥81	234,550 (8.9)	3728 (1.4)	238,278 (8.2)
Cohabitation status				
	Single	942,003 (35.9)	70,266 (25.5)	1,012,269 (34.9)
	Cohabiting	488,499 (18.6)	69,928 (25.4)	558,427 (19.3)
	Married	1,194,861 (45.5)	135,009 (49.1)	1,329,870 (45.8)
Comorbidities (n)				
	None	1,946,569 (74.1)	241,173 (87.6)	2,187,742 (75.4)
	1	426,677 (16.3)	27,329 (9.9)	454,006 (15.7)
	2	155,654 (5.9)	4647 (1.7)	160,301 (5.5)
	≥3	96,463 (3.7)	2054 (0.8)	98,517 (3.4)
Education (years)				
	<10	601,787 (22.9)	42,890 (15.6)	644,677 (22.2)
	10-15	1,175,364 (44.8)	124,310 (45.2)	1,299,674 (44.8)
	>15	807,635 (30.8)	105,160 (38.2)	912,795 (31.5)
	Unknown	40,577 (1.5)	2843 (1)	43,420 (1.5)
Ethnicity				
	Non-Western	195,638 (7.4)	21,790 (7.9)	217,428 (7.5)
	Western, not born in Denmark	54,761 (2.1)	5773 (2.1)	60,534 (2.1)
	Native, born in Denmark	2,325,363 (90.5)	247,640 (90)	2,622,604 (90.4)
Income (quintiles)				
	1	487,441 (18.6)	50,997 (18.5)	538,438 (18.6)
	2	598,552 (22.8)	45,088 (16.4)	643,640 (22.2)
	3	586,618 (22.3)	62,783 (22.8)	649,401 (22.4)
	4	540,238 (20.6)	65,139 (23.7)	605,377 (20.9)
	5	405,990 (15.5)	50,674 (18.4)	456,664 (15.7)
	Negative or zero	6524 (0.2)	522 (0.2)	7046 (0.2)
Urbanization (population)				
	>100,000	507,926 (19.4)	54,166 (19.7)	562,092 (19.4)
	20,000-100,000	655,472 (25)	67,516 (24.5)	722,988 (24.9)
	1,000-20,000	867,341 (33)	85,347 (31)	952,688 (32.9)
	<1,000	593,691 (22.6)	68,121 (24.8)	661,812 (22.8)

Characteristics		Telephone contacts (n=2,625,363, 90.5%), n (%)	Video contacts (n=275,203, 9.5%), n (%)	Total (n=2,900,566, 100%), n (%)
	Unplaceable	933 (0)	53 (0)	986 (0)
Employment status				
	Unemployed	332,428 (12.7)	27,305 (9.9)	359,733 (12.4)
	Retired	252,098 (20)	11,941 (4.3)	537,039 (18.5)
	Employed	1,767,837 (67.3)	235,957 (85.8)	2,003,794 (69.1)
Time of contact (hours until opening time of daytime general practice)				
	>4 before office hours	2,527,937 (96.3)	268,668 (97.6)	2,796,605 (96.4)
	<4 before office hours	97,426 (3.7)	6535 (2.4)	103,961 (3.6)
Region				
	North Denmark Region	405,869 (15.5)	34,556 (12.5)	440,415 (15.2)
	Central Denmark Region	989,549 (37.7)	88,821 (32.3)	1,078,370 (37.2)
	Region of Southern Denmark	946,931 (36.1)	134,029 (48.7)	1,080,960 (37.3)
	Region Zealand	283,014 (10.7)	17,807 (6.5)	300,821 (10.4)

Proportion of Video Contacts

During the study period, 9.5% (275,203/2,900,566) of telephone triage contacts to OOH-PC were video contacts. After the introduction of video, a range of 5%-15% video contacts was achieved within weeks across all regions. This level remained stable throughout the study period (data not shown).

Association Between Video Contact and Patient- and Contact-Related Characteristics

The frequency of video contacts was unevenly distributed across patient- and contact-related characteristics. The strongest associations were seen for age, comorbidity, employment status, region, and time of contact (Figure 1). Patients aged <20 years had a notably higher frequency of video contacts than patients aged 21 to 40 years (aIRR range: 2.39-1.31). This was also the case for employed compared to unemployed patients (aIRR 1.21). The frequency of video contacts was significantly higher for contacts to OOH-PC at more than 4 hours before the opening of daytime general practice (aIRR 1.40; reference: ≤4 hours), and was more frequent in the Region of Southern Denmark (aIRR 1.55; reference: North Denmark Region). In contrast, the frequency of video contacts was significantly lower for patients >40 years (aIRR range: 0.40-0.90; reference: 21-40 years), patients with comorbidities (aIRR range 0.59-0.90; reference: no comorbidities), and retired patients (aIRR 0.66; reference: unemployed). The frequency of video contacts was also significantly lower in Region Zealand (aIRR 0.73; reference: North Denmark Region).

Triage Outcomes

Patients receiving a video contact had a significantly higher frequency of ending the contact with advice and self-care compared to patients receiving a telephone contact (aIRR 1.21, 95% CI 1.21-1.21) (Table 2). Conversely, patients receiving a video contact had a significant lower frequency of being referred to a clinic consultation (aIRR 0.59, 95% CI 0.59-0.60) or a home visit compared to patients receiving a telephone contact (aIRR 0.31, 95% CI 0.29-0.32). The frequency of being admitted to a hospital was significantly higher after a video contact compared to a telephone contact (aIRR 1.20, 95% CI 1.17-1.23).

Table . Frequency of triage outcomes and their association with video contacts (incidence rate ratio).

Outcome	Telephone contacts (n=2,625,363), n (%)	Video contacts (n=275,203), n (%)	Total (n=2,900,566), n (%)	Incidence rate ratio (95% CI)	
				Crude	Adjusted[a]
Advice and self-care	1,663,681 (63.4)	215,484 (78.3)	1,879,567 (64.8)	1.24 (1.23-1.24)	1.21 (1.21-1.21)
Clinic consultation	712,255 (27.1)	49,262 (17.9)	759,948 (26.2)	0.66 (0.66-0.67)	0.59 (0.59-0.60)
Home visit	165,052 (6.3)	1926 (0.7)	168,233 (5.8)	0.12 (0.11-0.12)	0.31 (0.29-0.32)
Hospital admission	84,375 (3.2)	8531 (3.1)	92,818 (3.2)	0.95 (0.93-0.97)	1.20 (1.17-1.23)

[a]Adjusted for patient sex, age, cohabitation status, comorbidity, educational level, ethnicity, income, urbanization, employment status, region, and time of contact.

Follow-Up Contacts

In general, patients receiving a video contact had a significantly higher frequency of no follow-up contact compared to patients receiving a telephone contact (aIRR 1.09, 95% CI 1.08-1.09) (Table 3). For those who had a follow-up contact, the patients who received a video contact had a significantly higher frequency of having a follow-up contact with their regular GP compared to those receiving a telephone contact (aIRR 1.02, 95% CI 1.01-1.03). Conversely, patients receiving a video contact had a significant lower frequency of a follow-up contact in OOH-PC (aIRR 0.96, 95% CI 0.95-0.97) or at the hospital (aIRR 0.75, 95% CI 0.74-0.76) compared to patients receiving a telephone contact.

Table . Frequency of follow-up contacts and association between use of video contacts and subsequent follow-up contacts (incidence rate ratio).

Type of follow-up contact	Telephone contacts (n=2,625,363), n (%)	Video contacts (n=275,203), n (%)	Total (n=2,900,566)	Incidence rate ratio (95% CI)	
				Crude	Adjusted[a]
No follow-up	1,097,402 (41.8)	137,601 (50)	1,232,741 (42.5)	1.20 (1.19-1.20)	1.09 (1.08-1.09)
Daytime general practice[b]	719,349 (27.4)	70,728 (25.7)	791,854 (27.3)	0.94 (0.93-0.94)	1.02 (1.01-1.03)
OOH-PC[c,d]	396,430 (15.1)	37,703 (13.7)	435,085 (15)	0.90 (0.90-0.91)	0.96 (0.95-0.97)
Hospital[e]	412,182 (15.7)	29,171 (10.6)	440,886 (15.2)	0.68 (0.67-0.69)	0.75 (0.74-0.76)

[a]Adjusted for patient's sex and age, cohabitation status, comorbidity, educational level, ethnicity, income, urbanization, employment status, region, and time of contact.

[b]Contacts (telephone contacts, video contacts, clinic consultations, or home visits) to daytime general practice within 7 days from the index contact to OOH-PC.

[c]OOH-PC: out-of-hours primary care.

[d]All telephone triage contacts to OOH-PC within 7 days from the index contact to OOH-PC.

[e]All nonscheduled hospital contacts (emergency department visits and hospital admissions) within 1 day from the index contact to OOH-PC.

Discussion

Principal Results

Video was used in 9.5% (275,203/2,900,566) of all telephone triage contacts to OOH-PC. Video contacts were unevenly distributed across patient- and contact-related characteristics; video contacts were more often used for patients who were employed, young, without comorbidities, and contacting OOH-PC more than 4 hours before the opening hours of daytime general practice. Compared to telephone contacts, significantly more video contacts ended with advice and self-care and significantly fewer had follow-up contacts.

Strengths and Limitations

This study was based on a large data set, including codes for remuneration by GPs. The economic incentive for GPs to register all services provided contributed to the completeness of the data, though validity has not been studied [22].

Our study also had some limitations. First, we had no information on the reasons for encounters (RFEs), as this is not systematically registered in OOH-PC contacts. In each telephone triage contact, the triage GP assessed the relevance of video use based on the current RFE balanced against the specific patient- and contact-related characteristics. Therefore, telephone contacts and video contacts had different diagnostic scope, which could have influenced the differences found in triage outcome and follow-up contacts through confounding by indication. Second, we followed each patient for 7 days to record follow-up contacts to OOH-PC and to daytime general practice, as previously described in the literature [26]. This led to an overestimation of follow-up contacts, as we could not link these follow-up contacts to the index contact in OOH-PC using the RFE. However, any overestimation would be independent of type of contact. Finally, we used the Charlson Comorbidity Index to define comorbidity based on hospital diagnosis codes. This approach might have led to an underestimation of comorbidity [27], as patients with mild chronic diseases are often treated solely in general practice.

Several factors must be considered when generalizing the results of this study. First, the study period was defined according to the date of initiation of video contact in each of the regions. Therefore, the regions were included in different periods of the COVID-19 pandemic, and they had different contact patterns to primary care both inside and outside office hours [8,12,20,28] and probably also different distributions of triage outcomes. Second, triage GPs perform telephone triage with no decision support tool in Danish OOH-PC; this is unlike most countries with comparable OOH-PC services, which often use other health care professionals with decision support systems [3]. Compared to other triage professionals, GPs may be able to triage more patients via video contact. Lastly, Danish triage GPs were remunerated on a fee-for-service basis. As the fee for a video contact was higher than the fee for a telephone contact, this could have been an incentive to aim for a higher share of video contacts in this setting compared to countries with other payment structures.

Comparison With Prior Work

We found a 9.5% rate of use of video contacts to OOH-PC. To our knowledge, no previous studies used a data collection period of this length to report on video use in OOH-PC. Studies on

changing contact patterns in OOH-PC during the COVID-19 pandemic have found an overall increase in telehealth consultations (email, video, or telephone) [8,12,28]. Video use in daytime general practice has previously been reported to range from 1% to 6.4% [15,29-32]. However, as patient populations and RFEs are known to differ between daytime general practice and OOH-PC [33], these results cannot be compared with our findings. Furthermore, video contacts in OOH-PC guide triage professionals in the assessment of patients and in improving patient reassurance [13,14]. In contrast, video contact has often been used as a substitute for clinic consultations in daytime general practice for practical reasons, for example, to reduce travel time or limit the risk of contamination, but both patients and GPs seem to prefer in-person consultations in the postpandemic era [10,13,18].

Our study showed that video contacts were used more often for employed young patients without comorbidities. To the best of our knowledge, this is the first study to report on associations between patient- and contact-related characteristics and video contacts in OOH-PC. Studies conducted in daytime general practice have found higher video rates of use during COVID-19 lockdown periods [34] and among people from socioeconomically advantaged areas [34,35]. Previous studies have reported inconsistent results on the association between patient age and video use, as higher use has been reported for both younger [32,35] and older patients [30]. Moreover, daytime video use seems to be associated with patients with high morbidity [36] compared to patients with low morbidity. These findings are not in line with our study results, which could be due to differences in patient populations between daytime general practice and OOH-PC [33]. Furthermore, some previous studies were conducted during the peak of the COVID-19 pandemic, and different countries have different health care systems and had different approaches to tackling the pandemic.

We found that video contacts more often ended with advice and self-care and no follow-up contact compared to telephone contacts. Two qualitative studies investigating the effect of video contacts on the patient flow in daytime general practice found that GPs experience uncertainties when referring patients to secondary care after a video contact [9,37]. A UK study on follow-up contact after using a video contact service (used by hospitals, daytime general practices, and other services) found no significant difference in the number of subsequent referrals compared to telephone contacts [38]. However, these studies focused on video use in the daytime rather than on telephone triage in OOH-PC.

Implications for Practice and Future Research

Our study suggests that video contacts could help reduce the use of health care resources in the OOH-PC setting by lowering the number of subsequent clinic consultations and home visits. More studies are needed on the effect of video contact on patient flow. First, further research is needed to investigate the impact of video contact in relation to different RFEs. Second, future studies should explore if the findings of this study are maintained in the postpandemic period and across different OOH-PC organizations. Third, future studies should investigate if the video option might generate more contacts to OOH-PC overall. Fourth, our study indicates an association between video contacts and specific patient characteristics: video was more often used for employed young patients without comorbidities. This finding contrasts with most studies in daytime general practice and should be further investigated. Finally, it is important to note that we did not study costs associated with video use and its effects on resource use. Therefore, future studies should investigate the costs as well.

Conclusion

This study supports our hypothesis that video contacts could reduce use of health care resources in OOH-PC. Video use lowered the frequency of referrals to a clinic consultation or home visit and also lowered the frequency of follow-up contacts. However, the results could be biased due to confounding by indication, reflecting that triage GPs use video for a specific set of RFEs.

Acknowledgments

We gratefully acknowledge the financial support for this study provided by the Danish Health Insurance Foundation (Sygeforsikringen danmark), the General Practice Research Foundation of the Central Denmark Region (Praksisforskningsfonden), and the Department of Public Health, Aarhus University. The funding bodies had no role in the study design, data collection, data analysis, data interpretation, writing of the manuscript, or submission of the final article.

Authors' Contributions

All authors contributed to the study design, interpretation of results, and drafting and revising of the manuscript. MAN and CHV conducted the data management and the statistical analysis. All authors have agreed to the final submitted version of the manuscript.

Conflicts of Interest

None declared.

References

1. Pedersen KM, Andersen JS, Søndergaard J. General practice and primary health care in Denmark. J Am Board Fam Med 2012 Mar;25 Suppl 1:S34-S38. [doi: 10.3122/jabfm.2012.02.110216] [Medline: 22403249]

2. Smits M, Rutten M, Keizer E, Wensing M, Westert G, Giesen P. The development and performance of after-hours primary care in the Netherlands: a narrative review. Ann Intern Med 2017 May 16;166(10):737-742. [doi: 10.7326/M16-2776] [Medline: 28418455]
3. Steeman L, Uijen M, Plat E, Huibers L, Smits M, Giesen P. Out-of-hours primary care in 26 European countries: an overview of organizational models. Fam Pract 2020 Nov 28;37(6):744-750. [doi: 10.1093/fampra/cmaa064] [Medline: 32597962]
4. Drerup B, Espenschied J, Wiedemer J, Hamilton L. Reduced no-show rates and sustained patient satisfaction of telehealth during the COVID-19 pandemic. Telemed J E Health 2021 Dec;27(12):1409-1415. [doi: 10.1089/tmj.2021.0002] [Medline: 33661708]
5. Green MA, McKee M, Katikireddi SV. Remote general practitioner consultations during COVID-19. Lancet Digit Health 2022 Jan;4(1):e7. [doi: 10.1016/S2589-7500(21)00279-X] [Medline: 34952678]
6. Johnsen TM, Norberg BL, Kristiansen E, et al. Suitability of video consultations during the COVID-19 pandemic lockdown: cross-sectional survey among Norwegian general practitioners. J Med Internet Res 2021 Feb 8;23(2):e26433. [doi: 10.2196/26433] [Medline: 33465037]
7. Saint-Lary O, Gautier S, Le Breton J, et al. How GPs adapted their practices and organisations at the beginning of COVID-19 outbreak: a French national observational survey. BMJ Open 2020 Dec 2;10(12):e042119. [doi: 10.1136/bmjopen-2020-042119] [Medline: 33268433]
8. Sigurdsson EL, Blondal AB, Jonsson JS, et al. How primary healthcare in Iceland swiftly changed its strategy in response to the COVID-19 pandemic. BMJ Open 2020 Dec 7;10(12):e043151. [doi: 10.1136/bmjopen-2020-043151] [Medline: 33293329]
9. Wherton J, Greenhalgh T, Shaw SE. Expanding video consultation services at pace and scale in Scotland during the COVID-19 pandemic: national mixed methods case study. J Med Internet Res 2021 Oct 7;23(10):e31374. [doi: 10.2196/31374] [Medline: 34516389]
10. Due TD, Thorsen T, Andersen JH. Use of alternative consultation forms in Danish general practice in the initial phase of the COVID-19 pandemic - a qualitative study. BMC Fam Pract 2021 Jun 2;22(1):108. [doi: 10.1186/s12875-021-01468-y] [Medline: 34078281]
11. Joy M, McGagh D, Jones N, et al. Reorganisation of primary care for older adults during COVID-19: a cross-sectional database study in the UK. Br J Gen Pract 2020 Aug;70(697):e540-e547. [doi: 10.3399/bjgp20X710933] [Medline: 32661009]
12. Ramerman L, Rijpkema C, Bos N, Flinterman LE, Verheij RA. The use of out-of-hours primary care during the first year of the COVID-19 pandemic. BMC Health Serv Res 2022 May 21;22(1):679. [doi: 10.1186/s12913-022-08096-x] [Medline: 35597939]
13. Greenhalgh T, Ladds E, Hughes G, et al. Why do GPs rarely do video consultations? Qualitative study in UK general practice. Br J Gen Pract 2022 May;72(718):e351-e360. [doi: 10.3399/BJGP.2021.0658] [Medline: 35256385]
14. Gren C, Egerod I, Linderoth G, et al. “We can't do without it”: parent and call-handler experiences of video triage of children at a medical helpline. PLoS One 2022;17(4):e0266007. [doi: 10.1371/journal.pone.0266007] [Medline: 35421109]
15. Assing Hvidt E, Christensen NP, Grønning A, Jepsen C, Lüchau EC. What are patients’ first-time experiences with video consulting? A qualitative interview study in Danish general practice in times of COVID-19. BMJ Open 2022 Apr 15;12(4):e054415. [doi: 10.1136/bmjopen-2021-054415] [Medline: 35428624]
16. Mold F, Hendy J, Lai YL, de Lusignan S. Electronic consultation in primary care between providers and patients: systematic review. JMIR Med Inform 2019 Dec 3;7(4):e13042. [doi: 10.2196/13042] [Medline: 31793888]
17. Koch S, Guhres M. Physicians’ experiences of patient-initiated online consultations in primary care using direct-to-consumer technology. Stud Health Technol Inform 2020 Jun 16;270:643-647. [doi: 10.3233/SHTI200239] [Medline: 32570462]
18. Meurs M, Keuper J, Sankatsing V, Batenburg R, van Tuyl L. “Get used to the fact that some of the care is really going to take place in a different way”: general practitioners' experiences with e-health during the COVID-19 pandemic. Int J Environ Res Public Health 2022 Apr 22;19(9):5120. [doi: 10.3390/ijerph19095120] [Medline: 35564519]
19. Nordtug M, Assing Hvidt E, Lüchau EC, Grønning A. General practitioners’ experiences of professional uncertainties emerging from the introduction of video consultations in general practice: qualitative study. JMIR Form Res 2022 Jun 14;6(6):e36289. [doi: 10.2196/36289] [Medline: 35653607]
20. Huibers L, Bech BH, Kirk UB, Kallestrup P, Vestergaard CH, Christensen MB. Contacts in general practice during the COVID-19 pandemic: a register-based study. Br J Gen Pract 2022 Nov;72(724):e799-e808. [doi: 10.3399/BJGP.2021.0703] [Medline: 36253113]
21. Pedersen CB. The Danish civil registration system. Scand J Public Health 2011 Jul;39(7 Suppl):22-25. [doi: 10.1177/1403494810387965] [Medline: 21775345]
22. Andersen JS, Olivarius NDF, Krasnik A. The Danish national health service register. Scand J Public Health 2011 Jul;39(7 Suppl):34-37. [doi: 10.1177/1403494810394718] [Medline: 21775348]
23. 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;7:449-490. [doi: 10.2147/CLEP.S91125] [Medline: 26604824]

24. Thygesen LC, Daasnes C, Thaulow I, Brønnum-Hansen H. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. Scand J Public Health 2011 Jul;39(7 Suppl):12-16. [doi: 10.1177/1403494811399956] [Medline: 21898916]
25. Nguyen AD, Frensham LJ, Baysari MT, Carland JE, Day RO. Patients' use of mobile health applications: what general practitioners think. Fam Pract 2019 Mar 20;36(2):214-218. [doi: 10.1093/fampra/cmy052] [Medline: 29873708]
26. van Uden CJT, Zwietering PJ, Hobma SO, et al. Follow-up care by patient's own general practitioner after contact with out-of-hours care. A descriptive study. BMC Fam Pract 2005 Jun 9;6(1):23. [doi: 10.1186/1471-2296-6-23] [Medline: 15946382]
27. Prior A, Fenger-Grøn M, Larsen KK, et al. The association between perceived stress and mortality among people with multimorbidity: a prospective population-based cohort study. Am J Epidemiol 2016 Aug 1;184(3):199-210. [doi: 10.1093/aje/kwv324] [Medline: 27407085]
28. Morreel S, Philips H, Verhoeven V. Organisation and characteristics of out-of-hours primary care during a COVID-19 outbreak: a real-time observational study. PLoS One 2020;15(8):e0237629. [doi: 10.1371/journal.pone.0237629] [Medline: 32790804]
29. Scott A, Bai T, Zhang Y. Association between telehealth use and general practitioner characteristics during COVID-19: findings from a nationally representative survey of Australian doctors. BMJ Open 2021 Mar 24;11(3):e046857. [doi: 10.1136/bmjopen-2020-046857] [Medline: 33762248]
30. Murphy M, Scott LJ, Salisbury C, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021 Feb;71(704):e166-e177. [doi: 10.3399/BJGP.2020.0948] [Medline: 33558332]
31. Chang JE, Lindenfeld Z, Albert SL, et al. Telephone vs. video visits during COVID-19: safety-net provider perspectives. J Am Board Fam Med 2021;34(6):1103-1114. [doi: 10.3122/jabfm.2021.06.210186] [Medline: 34772766]
32. Dai Z, Sezgin G, Hardie RA, et al. Sociodemographic determinants of telehealth utilisation in general practice during the COVID-19 pandemic in Australia. Intern Med J 2023 Mar;53(3):422-425. [doi: 10.1111/imj.16006] [Medline: 36624629]
33. Huibers L, Moth G, Bondevik GT, et al. Diagnostic scope in out-of-hours primary care services in eight European countries: an observational study. BMC Fam Pract 2011 May 13;12:30. [doi: 10.1186/1471-2296-12-30] [Medline: 21569483]
34. Savira F, Orellana L, Hensher M, et al. Use of general practitioner telehealth services during the COVID-19 pandemic in regional Victoria, Australia: retrospective analysis. J Med Internet Res 2023 Feb 7;25:e39384. [doi: 10.2196/39384] [Medline: 36649230]
35. Rodriguez JA, Betancourt JR, Sequist TD, Ganguli I. Differences in the use of telephone and video telemedicine visits during the COVID-19 pandemic. Am J Manag Care 2021 Jan;27(1):21-26. [doi: 10.37765/ajmc.2021.88573] [Medline: 33471458]
36. Glazier RH, Green ME, Wu FC, Frymire E, Kopp A, Kiran T. Shifts in office and virtual primary care during the early COVID-19 pandemic in Ontario, Canada. CMAJ 2021 Feb 8;193(6):E200-E210. [doi: 10.1503/cmaj.202303] [Medline: 33558406]
37. Randhawa RS, Chandan JS, Thomas T, Singh S. An exploration of the attitudes and views of general practitioners on the use of video consultations in a primary healthcare setting: a qualitative pilot study. Prim Health Care Res Dev 2019 Jan;20:e5. [doi: 10.1017/S1463423618000361] [Medline: 29909798]
38. Smith C, Kubanova B, Ahmed F, Manickavasagam J. The effectiveness of remote consultations during the COVID-19 pandemic: a tool for modernising the national health service (NHS). Cureus 2022 Dec;14(12):e32301. [doi: 10.7759/cureus.32301] [Medline: 36627990]

Abbreviations

aIRR: adjusted incidence rate ratio
GDPR: General Data Protection Regulation
GP: general practitioner
ICPC-2: International Classification of Primary Care, 2nd Edition
IRR: incidence rate ratio
OOH-PC: out-of-hours primary care
RFE: reason for encounter
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by C Lovis; submitted 10.03.23; peer-reviewed by F Savira, T Koskela; revised version received 07.02.24; accepted 10.02.24; published 04.04.24.

Please cite as:
Nebsbjerg MA, Vestergaard CH, Bomholt KB, Christensen MB, Huibers L
Use of Video in Telephone Triage in Out-of-Hours Primary Care: Register-Based Study
JMIR Med Inform 2024;12:e47039
URL: https://medinform.jmir.org/2024/1/e47039
doi:10.2196/47039

Reducing Firearm Access for Suicide Prevention: Implementation Evaluation of the Web-Based "Lock to Live" Decision Aid in Routine Health Care Encounters

Julie Angerhofer Richards[1,2], MPH, PhD; Elena Kuo[1], MPH, PhD; Christine Stewart[1], PhD; Lisa Shulman[1], MSW; Rebecca Parrish[3], LICSW; Ursula Whiteside[4,5], PhD; Jennifer M Boggs[6], MSW, PhD; Gregory E Simon[1,3,5], MD, MPH; Ali Rowhani-Rahbar[7,8], MD, MPH, PhD; Marian E Betz[9], MD, MPH

1

2

3

4

5

6

7

8

9

Corresponding Author:
Julie Angerhofer Richards, MPH, PhD

Abstract

Background: "Lock to Live" (L2L) is a novel web-based decision aid for helping people at risk of suicide reduce access to firearms. Researchers have demonstrated that L2L is feasible to use and acceptable to patients, but little is known about how to implement L2L during web-based mental health care and in-person contact with clinicians.

Objective: The goal of this project was to support the implementation and evaluation of L2L during routine primary care and mental health specialty web-based and in-person encounters.

Methods: The L2L implementation and evaluation took place at Kaiser Permanente Washington (KPWA)—a large, regional, nonprofit health care system. Three dimensions from the RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) model—*Reach*, *Adoption*, and *Implementation*—were selected to inform and evaluate the implementation of L2L at KPWA (January 1, 2020, to December 31, 2021). Electronic health record (EHR) data were used to purposefully recruit adult patients, including firearm owners and patients reporting suicidality, to participate in semistructured interviews. Interview themes were used to facilitate L2L implementation and inform subsequent semistructured interviews with clinicians responsible for suicide risk mitigation. Audio-recorded interviews were conducted via the web, transcribed, and coded, using a rapid qualitative inquiry approach. A descriptive analysis of EHR data was performed to summarize L2L reach and adoption among patients identified at high risk of suicide.

Results: The initial implementation consisted of updates for clinicians to add a URL and QR code referencing L2L to the safety planning EHR templates. Recommendations about introducing L2L were subsequently derived from the thematic analysis of semistructured interviews with patients (n=36), which included (1) "have an open conversation," (2) "validate their situation," (3) "share what to expect," (4) "make it accessible and memorable," and (5) "walk through the tool." Clinicians' interviews (n=30) showed a strong preference to have L2L included by default in the EHR-based safety planning template (in contrast to adding it manually). During the 2-year observation period, 2739 patients reported prior-month suicide attempt planning or intent and had a documented safety plan during the study period, including 745 (27.2%) who also received L2L. Over four 6-month subperiods of the observation period, L2L adoption rates increased substantially from 2% to 29% among primary care clinicians and from <1% to 48% among mental health clinicians.

Conclusions: Understanding the value of L2L from users' perspectives was essential for facilitating implementation and increasing patient reach and clinician adoption. Incorporating L2L into the existing system-level, EHR-based safety plan template reduced the effort to use L2L and was likely the most impactful implementation strategy. As rising suicide rates galvanize the urgency of prevention, the findings from this project, including L2L implementation tools and strategies, will support efforts to promote safety for suicide prevention in health care nationwide.

(JMIR Med Inform 2024;12:e48007) doi:10.2196/48007

KEYWORDS

suicide prevention; firearm; internet; implementation; suicide; prevention; decision aid; risk; feasible; support; evaluation; mental health; electronic health record; tool

Introduction

Firearm-related suicide accounts for approximately half of the suicide deaths in the United States annually [1]. Firearms are common in Americans' lives [2]; about one-third of Americans report owning firearms [3], and an additional 10% report living in a household with a firearm [4], with higher rates in western states [2], among veterans [5], and in rural areas [6]. Moreover, the rate of ownership of new firearms appears to have increased recently among women, Black people, and Hispanic people [7]. Suicide attempts by firearm are highly lethal; researchers estimate that 85% to 95% of individuals who attempt suicide by firearm do not survive [8,9], and people with access to firearms, particularly if firearms are kept loaded and unlocked [10,11], have increased suicide risk [12,13]. Clinicians may have opportunities to intervene with patients at risk for firearm-related suicide because about 50% of individuals who die by suicide see a clinician in the month before death, and over 80% see one in the year before death [14]. Moreover, clinician-initiated discussions about reducing access to firearms have demonstrated effectiveness for improving firearm security practices (particularly in combination with free safe storage devices) [15-17], as well as promising findings for reducing suicide attempts [18,19].

Despite its potential benefits, clinician-initiated dialogue about limiting access to firearms is an uncommon practice across many primary care and mental health specialty practices [18,20]. Common barriers include time, clinicians' unfamiliarity with firearms, and concerns about negatively impacting relationships or alienating patients [21]. "Lock to Live" (L2L) is a self-directed, anonymous, web-based decision aid that was designed to address these barriers. L2L was developed in collaboration with clinicians, firearm owners, and people who had experienced suicidal thoughts and attempts [22]. Consistent with international design standards [23,24], the L2L decision aid steps users through various considerations regarding in-home and out-of-home firearm storage options, such as types of storage, costs of storage, and background check requirements, with a goal of encouraging storage solution discussions that are consistent with the users' values and preferences [22]. Two subsequent research studies demonstrated promising results for the feasibility and acceptability of offering L2L emergency care encounters [25] and for the uptake of L2L when it was offered via secure patient portal messages after outpatient care encounters [26]. Though L2L appears to be a useful tool for supporting suicide prevention in clinical practice, little is known about how to use L2L during routine health care encounters outside of a research context.

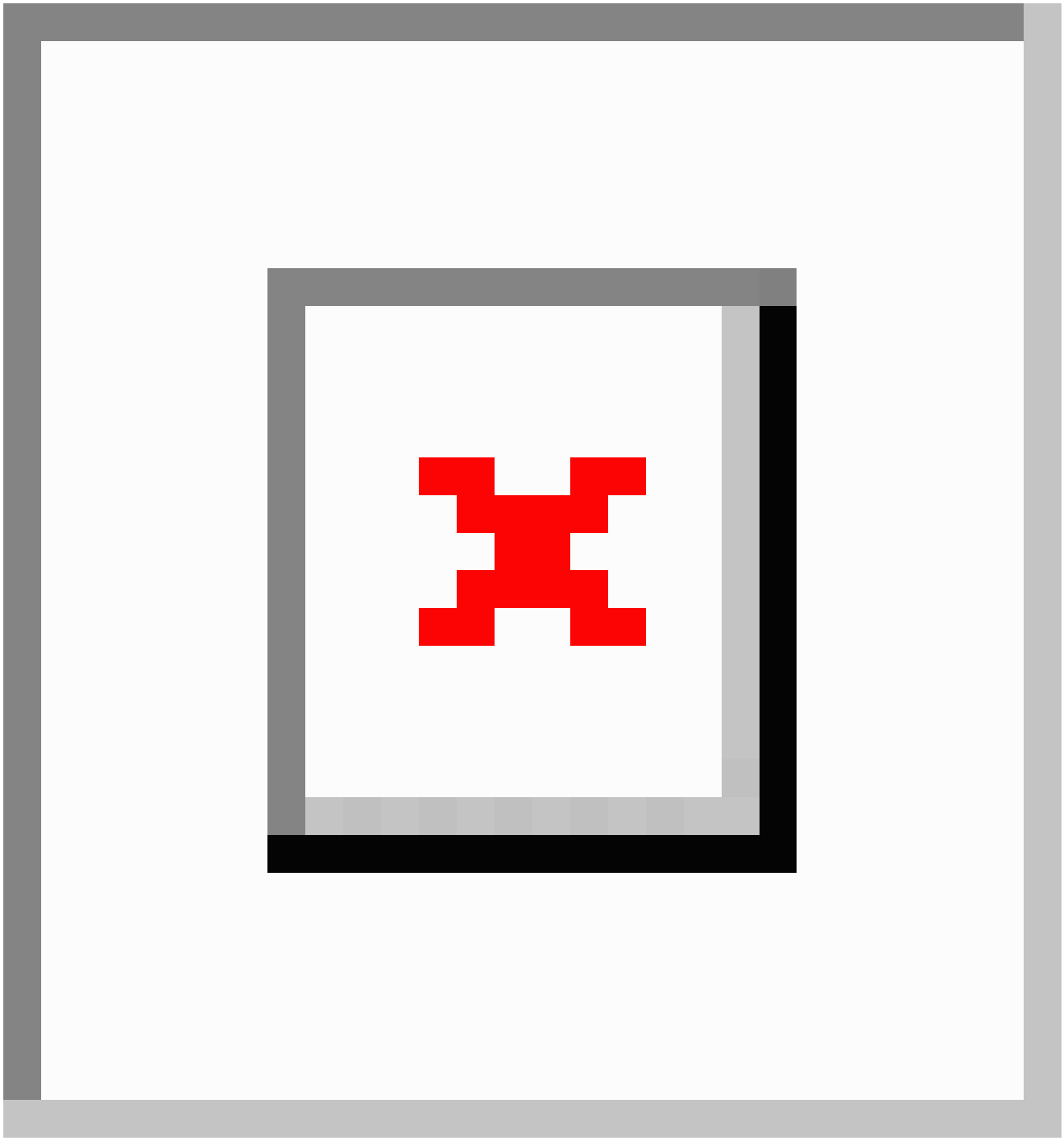
The goal of this project was to use mixed methods (qualitative and statistical evaluations) to support the implementation and evaluation of L2L during primary care and mental health specialty encounters in a large, regional health care system. Specifically, this project used semistructured interviews with clinicians and patients to support implementation, as well as statistical analyses to evaluate the reach and adoption of L2L over a 2-year period. The evaluation findings will inform considerations for implementing L2L nationwide to support suicide prevention in health care systems.

Methods

Setting

L2L implementation and evaluation took place at Kaiser Permanente Washington (KPWA)—1 of 8 regional Kaiser Permanente health care systems, which together form one of the nation's largest nonprofit health care organizations and serve 12.5 million people [27]. At the time of this evaluation, KPWA had provided comprehensive medical care to approximately 700,000 members across Washington State via employer-sponsored insurance plans, individual insurance plans, or capitated Medicaid or Medicare programs. In 2016, KPWA augmented standard clinical workflows to support the identification and engagement of patients at high risk of suicide attempts (Figure 1) [28,29]. Specifically, a system-level electronic health record (EHR) template was created to support clinician-initiated safety planning among patients who are identified as at high risk of suicide during primary care and mental health specialty encounters [28,30]. Nationally, safety planning is a widely recommended best practice [31], and KPWA had an established process for safety planning that included addressing access to lethal means but did not offer any specific resources to clinicians or patients about firearm storage options. Consistent with the goal of L2L, step 6 of this safety plan template was designed to support patients in limiting access to lethal means, such as firearms and prescription medications.

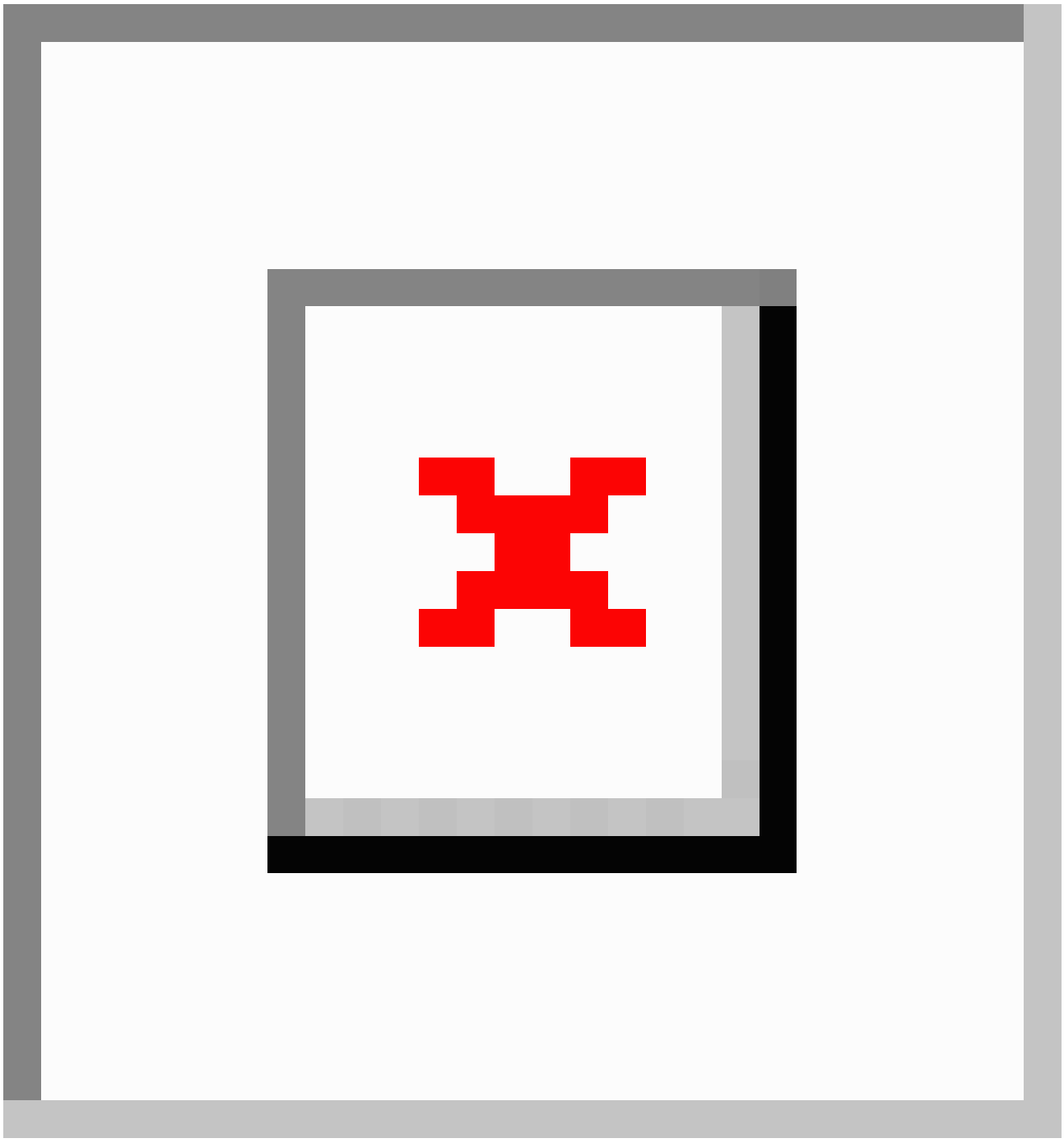
Figure 1. Clinical workflow for supporting the identification and engagement of patients at high risk of suicide during primary and mental health specialty encounters at Kaiser Permanente Washington.



Implementation and Evaluation Framework, Data Sources, and Study Design

Three dimensions from the RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) model—*Reach*, *Adoption*, and *Implementation* [32,33]—were selected to both inform and evaluate the implementation of L2L at KPWA (Multimedia Appendix 1) over a 2-year observation period (January 1, 2020, to December 31, 2021). Specifically, a qualitative, team-based, formative evaluation [34] was used, involving semistructured interviews with purposefully sampled patients and clinicians to facilitate implementation tools and strategies. Descriptive statistical analyses were used to evaluate the reach of L2L among patients identified at high risk of suicide and L2L adoption over a 2-year observation period. The findings were stratified by primary care and mental health specialty service settings due to the variation in the timing of the L2L trainings across these settings (Figure 2).

Figure 2. L2L implementation tools and strategies used over four 6-month subperiods of the 2-year observation period (January 1, 2020, to December 31, 2021). Tools are shown in blue, strategies are shown in green, and capped lines indicate semistructured interviews. EHR: electronic health record; KPWA: Kaiser Permanente Washington; L2L: Lock to Live; LICSW: licensed clinical social worker.



Semistructured Qualitative Interviews

EHR data were used to purposefully recruit adult patients to participate in semistructured interviews that included questions to elicit suggestions about introducing L2L, as part of a broader interview that focused on exploring perceptions regarding and experiences with firearm access assessment [35]. An invitation letter was mailed to sampled patients (age≥18 y) who had received a standardized question about firearm access ("Do you have access to guns? Yes/No") in the prior 2 weeks on a mental health questionnaire [30]. A stratified sampling distribution was used to recruit approximately equal numbers of patients in 3 groups, including those who (1) reported firearm access, (2) reported no firearm access, and (3) did not respond (ie, left the question blank). The three sampling groups were also designed to purposefully include patients who had reported thoughts about self-harm in the prior 2 weeks via the ninth question of the 9-item Patient Health Questionnaire (PHQ-9) [36]. Interviewers attempted to reach all invitees for 2 weeks to invite them to participate in a phone interview. Following the portion of the interview guide that focused on firearm access assessment [35], interviewers described how patients reporting suicidal thoughts would soon be receiving L2L and elicited feedback about how to introduce this tool in a way that would make patients more likely to use it (Multimedia Appendix 2). This portion of the interview transcript was extracted into Excel

(Microsoft Corporation) and analyzed by using a team-based, rapid, qualitative inquiry approach [37]. This involved an iterative data analysis wherein 2 coders independently coded the L2L portion of the transcript by using a combination of deductive and inductive content analyses, with codes developed a priori from the interview guide as well as codes that emerged from the interviews [38]. This was followed by several rounds of discussions with 2 additional team members who reviewed the coded data and reconciled themes iteratively for the purpose of using the summarized themes to facilitate implementation.

Following the completion of the qualitative analysis that focused on patient-informed L2L implementation, interviewers initiated clinician recruitment activities, which were also more broadly focused on firearm access assessment. At the recommendation of care delivery leaders, interviewers outreached to the following two groups of clinicians responsible for engaging patients identified at risk of suicide in risk mitigation (ie, safety planning [39]): (1) licensed clinical social workers (LICSWs) supporting integrated mental health in primary care [28] and (2) consulting nurses (registered nurses) responsible for connecting patients (ie, those reporting suicidality after business hours via telephone) to telephone-based follow-up care. The L2L portion of the interview guide included questions informed by patient interviews (Multimedia Appendix 2) and was analyzed by using the same rapid, qualitative inquiry approach [37] that was used for patient interviews for the purpose of further facilitating L2L implementation.

Descriptive Statistical Analyses

L2L Reach and Adoption

EHR data were used to summarize L2L "reach," which was defined as the proportion of patients identified as at high risk of suicide via routine screening and assessment clinical workflows (Figure 1) and received the web-based decision aid. Specifically, we described characteristics of patients who had a documented safety plan during the 2-year observation period and characteristics of patients who had a safety plan that included a reference to L2L. Next, we described the adoption of L2L by primary care and mental health specialty clinicians over four 6-month subperiods of the observation period by calculating the proportions of patients identified at high risk of suicide (via suicide risk assessment; described in the *Measures* section) who had a documented safety plan with a reference to L2L and those who had a documented safety plan without a reference to L2L. We selected 6-month subperiods as the most helpful way to visually describe L2L adoption over time, since implementation paused during the initial COVID-19 outbreak (described in the *Implementation Timeline, Tools, and Strategies* section). We stratified by service setting due the variation in the timing of L2L trainings for these groups of clinicians.

Measures

The Columbia Suicide Severity Rating Scale (C-SSRS) was used to measure suicide risk, as per current clinical workflows. Specifically, patients reporting some level of suicide attempt planning or intent in the past month (ie, answering "yes" to C-SSRS question 3 or higher) were considered to be at "high risk" and alerted clinicians (via EHR prompts) to initiate safety planning. Distinctive phrases from standard EHR-based templates were used to detect safety plans documented in the text of clinical notes among patients identified at high risk (Multimedia Appendix 3). Sociodemographic and clinical characteristics of interest, including those known to be associated with firearm ownership and suicide risk [40,41], were measured by using the following administrative and diagnostic EHR data: age (continuous); sex (male or female); race and ethnicity (Asian, Black, Hispanic or Latinx, White, other, or unknown); insurance type (commercial, Medicare, Medicaid, or other); rurality (urban, large suburban, small suburban, or mostly rural) [42,43]; and prior year mental health, substance use, and self-harm diagnoses derived from the *International Classification of Diseases, Tenth Revision, Clinical Modification*. Reported firearm access was measured based on a positive response to the question on the mental health questionnaire [30] that was also used for qualitative interview recruitment (described in the *Semistructured Qualitative Interviews* section).

Ethical Considerations

The project team received approval from the KPWA Region Institutional Review Board (review number: 1826198) to conduct this study. Patients who agreed to participate in the phone interview provided oral consent, including permission for the interview to be audio-recorded and professionally transcribed, and they received a US $50 cash incentive for participation. During clinician recruitment activities, clinicians received up to 3 email invitations, which included a study information sheet and instructions for opting out of participation and further contact. Participating clinicians verbally consented to participation and received a US $50 gift card for participation.

Results

Implementation Timeline, Tools, and Strategies

Over the 2-year implementation period (January 1, 2020, to December 31, 2021), a team of researchers and care delivery leaders took a pragmatic approach to iteratively creating and refining L2L implementation tools and strategies for primary care and mental health specialty clinicians (Figure 2). Initially, tools included an EHR-based macro (ie, EPIC SmartPhrase [Epic Systems Corporation]) for clinicians to easily add a URL and QR code referencing L2L to safety planning templates and a 1-page quick-start guide (ie, "Huddlecard") with information on how to use the new SmartPhrase during routine clinic meetings (ie, "huddles"). In February 2020, the LICSWs who supported mental health care delivery in primary care [28] received information about L2L during a brief, web-based staff training session. Additional trainings that were planned for mental health specialty clinicians were put on hold during the widespread service disruption that subsequently occurred in response to the initial COVID-19 outbreak in March 2020. Additional tools and strategies were used, following recommendations from the care delivery leaders responsible for primary care and mental health service recovery and from the patients and clinicians who participated in semistructured qualitative interviews (detailed in the *Findings From Semistructured Qualitative Interviews* section).

Findings From Semistructured Qualitative Interviews

Of 76 patients who were purposefully sampled during 2 waves of recruitment, 36 were interviewed from November 18, 2019, to February 10, 2020 (Table 1). Five organizing themes were derived from the portion of the interview that elicited perceptions and suggestions about L2L and were used to create a handout for clinicians, with suggestions about how to introduce L2L to their patients at risk of suicide (Multimedia Appendix 4), including recommendations to (1) "have an open conversation," (2) "validate their situation," (3) "share what to expect," (4) "make it accessible and memorable," and (5) "walk through the tool" (Table 2). In addition to these recommendations, patients expressed a preference for receiving information about L2L from "trusted" and "caring" clinicians.

Table . Characteristics of patient (n=36) and clinician (n=30) semistructured interview participants.

		Patients	Clinicians
Sex, n (%)			
	Female	17 (47)	24 (80)
	Male	19 (53)	6 (20)
Age (y), mean (SD)		47.3 (17.9)	44.3 (12.1)
Age category (y), n (%)			
	19-29	8 (22)	1 (3)
	30-49	11 (31)	20 (67)
	50-64	9 (25)	6 (20)
	≥65	8 (22)	3 (10)
Race and ethnicity, n (%)			
	American Indian or Alaska Native	0 (0)	1 (3)
	Black	3 (8)	2 (7)
	Asian or Pacific Islander	3 (8)	5 (17)
	Latinx or Hispanic	1 (3)	4 (13)
	Unknown	2 (6)	0 (0)
	White	27 (75)	18 (60)
Reported firearm access[a,b], n (%)		16 (44)	N/A[c]
Reported thoughts about self-harm (prior 2 wk)[a,d], n (%)		15 (42)	N/A

[a]Patients' responses recorded on the Kaiser Permanente Washington mental health monitoring questionnaire used for criterion sampling within the 2 wk prior to the recruitment initiation.

[b]"Do you have access to guns? Yes/No."

[c]N/A: not applicable.

[d]Ninth question on the 9-item Patient Health Questionnaire: "Thoughts that you would be better off dead, or of hurting yourself."

Table . Thematic analysis of semistructured interviews with patients (n=36) and recommendations for introducing "Lock to Live" (L2L).

Themes	Illustrative quotes[a]	Recommendation
Show caring and compassion, ask permission, and respect autonomy	• "I think it's important to just take a breath, sit down with them, hold their hand, look them in the eye - 'how can I help you? Help me help you. What's going on? Tell me. What are you thinking? How are you feeling? How can I help you?' Instead of an assembly line and 'I only have a few minutes,' so they [providers] don't take the time." (Patient B029) • "I would hope the provider is very warm and caring and explains it's a safety precaution, it's for your better health and it insures you'll be safer…basically it's another part of your little toolbox to keep yourself well." (Patient A032) • "Probably compassionately, potentially generalized to begin with, to find out if the person is resistant right upfront… Maybe you could put a question, 'would you be willing to consider options for storing or access to lethal means, whether it's firearms or medication? Is it something you would be willing to discuss and look into if you were experiencing suicidal thoughts?'" (Patient A005) • "We're offering you the means to protect yourself, this is not an us decision, this is a you decision. So here is the website, the online information and we encourage you to look at it, but it's your decision….Nobody can force you to do things. So bringing it up more as like – not we're taking it [firearm] away from you, but letting you decide what to do with it." (Patient A024) • "Reassuring people that their firearm ownership will not end because they're going through a rough patch in life; their ability to have their own authority to hold onto their possessions [will not end], firearms or not." (Patient B036)	"Have an open conversation": patients were more willing to listen and try a tool if a clinician took the time to connect, showed compassion for people's unique experiences, and showed respect for autonomy.
Frame as helpful resource and normalize experiences	• "I think overall education about the topic to start out with, just to say…'this is what we have found is helpful, in these situations.' Moving more into letting people know what their resources could be. Just education to be begin with, so it's not so threatening. I think anytime somebody's in a vulnerable place emotionally, they're already possibly feeling threatened and they may not want to trust a lot of people." (Patient A005) • "I would hope it would be pretty real, like a conversation…'based on your health concerns you're showing, we've got some important information we'd like to share with you,' especially if that person has a relationship or feels responsible with the person presenting it, would stay there together and talk about it afterwards….Also statistics to help a person realize how more common this is." (Patient A011)	"Validate their situation": normalize their experience, share how common suicidal thoughts are, and be nonjudgmental in your approach; people have a variety of gun beliefs.

Themes	Illustrative quotes[a]	Recommendation
Address privacy and security	• "[The provider] would have to explain what it does, how it's going to work and how private it is - nobody can get into your part of the website anyway, your personal page, where you go. So she has to reassure them about that….I just don't feel that being online is that secure." (Patient B004) • "privacy is probably number one and an assurance that you're not being turned in….I would be concerned in our surveillance state that disclosing things to a website about my firearm use might somehow come close to violating some kind of civil right to privacy." (Patient B008) • "As long as people don't have to put in information which can be tracked, I can see lots of people using it. I mean the minute [you have to enter] your name, address, phone number, medical ID number, whatever else, people are going to go – eh." (Patient A033)	"Share what to expect": address privacy and how information is stored if patients visit the website; assure patients that L2L is anonymous.
Accessibility is key	• "You don't want to make it hard to find on a website because it doesn't take me very long. If something's really hard to find on a website, I'm out of there." (Patient B004) • "If there are hoops to jump through before you can access it, if you have to log in, go through a bunch of pages - maybe if it was right there, ready to access at any time, I'd say that'd be better." (Patient B008) • "I'd love to see it everywhere. Have little cards that could be given out, a billboard, having my doctor [send it]." (Patient A033) • "If I knew it existed, I would probably try it. advertise it." (Patient A032) • "Highlight it in your After Visit Summary too." (Patient A002)	"Make it accessible and memorable": have multiple routes for sharing the website and sending reminders (after-visit summary, message, website, pamphlet).
Demonstrate and "show, don't just tell"	• "I think when you're in a pit of despair, to go and do it on your own, some people will do that and other people will not. They need to be taken by the hand and go, 'what do you think about this?' Read it together." (Patient B036) • "I'm more keen to follow somebody who's like 'I'm offering you the opportunity to maybe do this together,' instead of 'I'm watching out for you.'" (Patient A024) • "Being shown an example would be nice, showing it off briefly. Knowing more specifically what it does or how it could be helpful as opposed to just knowing it exists." (Patient B033) • "I think showing the patient or at least offering, would you like me to show you? Not just telling somebody, because short term memory is only like 30 s or a couple minutes and then you forget about it." (Patient A002)	"Walk through the tool": most patients said that a website walk-through, rather than simply having a conversation, would be helpful to overcome the barrier of trying something new, especially if already depressed.

[a]Identifier "A": patients in the first wave of interviews; identifier "B": patients in the second wave of interviews (grammatic edits, noted in brackets, were added to clarify intended meaning).

Of 51 purposefully sampled clinicians responsible for safety planning with patients identified at high risk of suicide, 30 were interviewed from July 7, 2020, to October 8, 2020 (Table 1), including 25 LICSWs and 5 registered nurses. During the

interviews with LICSWs, only 3 had actually used L2L with a patient—9 were unfamiliar with L2L, and 12 were familiar with but had not yet used L2L. Most clinicians saw clear benefits to L2L as an option for supporting both clinicians and patients. Several clinicians expressed concern about using the tool to replace dialogue about lethal means, and most supported the idea of a walk-through, as patients had recommended. Clinicians also expressed a strong preference to have L2L information included by default in the EHR-based safety planning template, in contrast to having clinicians remember to add it (via SmartPhrase). A clinician also suggested automatically including L2L in after-visit summaries when patients reported thoughts about self-harm on the PHQ-9. The implementation team worked with clinical partners to update the system-level, EHR-based safety plan template to include L2L information and updated the Huddlecard to communicate this change (Multimedia Appendix 5). After-visit summaries were used to provide safety plans to patients who were seen via a secure, web-based patient portal, and L2L was automatically included after the template change. Several clinicians also requested follow-up trainings or refreshers about L2L. The team therefore conducted a round of brief trainings, which were presented during routine clinic huddles with mental health specialty clinicians, and created a 3-minute training video (Multimedia Appendix 6).

Findings From Descriptive Statistical Analyses

During the study period, 2739 adult patients reported some prior-month suicide attempt planning or intent via routine suicide risk assessment workflows during primary care or mental health specialty encounters and had a documented safety plan, including 745 (27.2%) who also received L2L. Overall, there were no major differences in the demographic and clinical characteristics between patients who received L2L and the broader population that was identified as at risk of suicide and had a documented safety plan (Table 3).

Table . Characteristics of patients who received "Lock to Live" (L2L; n=745) and were among patients with a documented safety plan (n=2739) during the implementation period (January 1, 2020, to December 31, 2021).

		Patients who received L2L, n (%)	Patients with a documented safety plan[a], n (%)
Age[b] (y)			
	18-39	513 (68.9)	1817 (66.3)
	40-64	187 (25.1)	732 (26.7)
	≥65	45 (6)	190 (6.9)
Sex[c]			
	Female	445 (59.7)	1753 (64)
	Male	300 (40.3)	986 (36)
Race and ethnicity[c]			
	American Indian or Alaska Native	14 (1.9)	75 (2.7)
	Asian	54 (7.2)	199 (7.3)
	Black	52 (7)	166 (6.1)
	Hawaiian or Pacific Islander	8 (1.1)	45 (1.6)
	Hispanic or Latinx	59 (7.9)	201 (7.3)
	Unknown	93 (12.5)	294 (10.7)
	White	465 (62.4)	1759 (64.2)
Insurance[c]			
	Commercial	530 (71.1)	1831 (66.8)
	Medicare	83 (11.1)	342 (12.5)
	Medicaid	54 (7.2)	228 (8.3)
	Not enrolled	78 (10.5)	338 (12.3)
Rural or urban [c, d]			
	Urban	301 (40.4)	1010 (36.9)
	Large suburban	205 (27.5)	799 (29.2)
	Smaller suburban	186 (25)	802 (29.3)
	Mostly rural	31 (4.2)	96 (3.5)
Mental health diagnoses[e]			
	Depression	675 (90.6)	2502 (91.3)
	Anxiety	652 (87.5)	2434 (88.9)
	Serious mental illness	144 (19.3)	586 (21.4)
	Substance use disorder	196 (26.3)	752 (27.5)
	Suicide attempt diagnosis	39 (5.2)	175 (6.4)
Reported firearm access[e]		150 (20.1)	501 (18.3)

[a]Includes safety plans with L2L.

[b]At evaluation midpoint (January 1, 2021).

[c]At first encounter.

[d]Missing information for 22 patients.

[e]During implementation period.

The adoption of L2L increased substantially over the 2-year observation period (Tables 4 and 5). During this time, rates of documented safety plans among patients identified at high risk of suicide (C-SSRS score≥3) remained fairly consistent—51.2%

to 55.2% of primary care patients and 73.4% to 78.4% of mental health specialty patients had a documented safety plan. However, over four 6-month subperiods of the observation period, L2L adoption rates increased substantially from 2% to 29% among primary care clinicians and <1% to 48% among mental health clinicians, increasing primarily after L2L was integrated into the EHR-based safety planning template.

Table . Proportions of primary care patients who were identified as at high risk of suicide and had a documented safety plan during primary care encounters over the implementation period (January 1, 2020, to December 31, 2021).

Subperiods of implementation period	Patients with a documented safety plan that did not include "Lock to Live," %	Patients with a documented safety plan that did include "Lock to Live," %
Months 1-6	53.5	1.6
Months 7-12	50	4.1
Months 13-18	41.7	11.1
Months 19-24	22.2	29.1

Table . Proportions of primary care patients who were identified as at high risk of suicide and had a documented safety plan during mental health specialty encounters over the implementation period (January 1, 2020, to December 31, 2021).

Subperiods of implementation period	Patients with a documented safety plan that did not include "Lock to Live," %	Patients with a documented safety plan that did include "Lock to Live," %
Months 1-6	78.1	0.3
Months 7-12	74.8	1.4
Months 13-18	53.8	19.6
Months 19-24	25.9	48.4

Discussion

This novel study used mixed methods to support the implementation and evaluation of a web-based decision aid that was designed to help patients at risk of suicide limit access to firearms. Specifically, findings from semistructured interviews with patients and clinicians were used to facilitate L2L implementation, while statistical analyses were used to describe rates of reach among patients identified at risk of suicide and increased adoption by clinicians who cared for them during the 2-year observation period.

L2L development centered users' values and preferences in the design process [22]. Similarly, the tools and strategies developed for this project used information from semistructured interviews with people who were the most likely to be impacted by L2L implementation, including firearm owners, patients experiencing suicidality, and the clinicians who care for them. Clinicians have reported a lack of experience with handling firearms and have expressed apprehension about discussing firearm safety due to concerns about damaging relationships with patients [44-46]. Likewise, patients have expressed apprehension about disclosing access to firearms due to concerns about privacy, autonomy, and firearm ownership rights [47,48]. For these reasons, patients and clinicians perceive firearm access assessment as challenging but also as valuable for supporting suicide prevention [35]. This implementation project showed that clinicians, that is, those responsible for engaging at-risk primary care and mental health patients in suicide risk mitigation, willingly adopted the use of L2L to support safety planning.

This study also has important implementation implications. Unsurprisingly, the rates of L2L adoption increased after L2L was incorporated into the existing system-level safety planning template as a default (primarily in the latter half of year 2). This finding underscores the importance of removing barriers to the adoption of web-based decision aids and making adoption "easy" [49]. In contrast, those seeking change often focus on amplifying benefits or "selling" their new idea or innovation; however, it may be equally as important or more important to focus on "friction," that is, "psychological forces that oppose and undermine change," such as inertia, effort, emotion, and reactance [50]. In the case of L2L, reducing the effort required for clinicians to remember to use L2L appeared to be the main driver of its adoption. However, the tools and strategies that were designed to communicate about the benefits of using L2L (eg, training, video, Huddlecard, and newsletter information) were likely necessary for leaders to understand L2L's value to patients and clinicians and approve the system-level change that was required to make L2L easier to use for clinicians.

This study has important clinical implications for supporting suicide prevention in health care. First, L2L supports clinicians who engage patients identified at risk of suicide in collaborative safety planning and lethal means counseling, which are evidence-based suicide risk mitigation practices that are recommended by the Zero Suicide Institute and follow the principles outlined in the National Strategy for Suicide Prevention [29,51-53]. Moreover, the recommendations from interview participants ("have an open conversation," "share what to expect," and "walk through the tool") support a motivational interviewing approach to lethal means counseling and align with the recommendations of the Veterans Health Administration [54]. Second, L2L was developed by patients with lived experiences of suicidality and firearm ownership; therefore, L2L supports cultural competency in health care as

a culturally aligned intervention [55]. Finally, this technology-based, EHR-embedded approach to addressing lethal means supports all 6 aims of health care quality that are outlined by the Institute of Medicine—*safe*, *effective*, *patient-centered*, *timely*, *efficient*, and *equitable* [56,57].

There are several limitations of this project that have implications for future research. First, the implementation of L2L at KPWA occurred during the initial outbreak of a global pandemic, which impacted the original implementation plans while health care systems responded to the pandemic and rapidly shifted toward providing web-based mental health care [58]. Semistructured interviews with patients took place prior to this shift. Future research should explore optimizing mental health care delivery workflows that support web-based suicide risk identification (ie, screening and assessment) [59] and incorporating L2L in web-based care encounters via secure telehealth platforms that are designed to support patient engagement. Second, L2L recognizes and addresses firearm policies related to background checks and how these policies might influence the legality of temporary firearm transfers for addressing suicide risk, but it does not address specific state laws. Additional work to understand the legality of recommendations about firearm safety practices may be helpful for health care systems that implement L2L. Third, this project was not designed to measure the specific impact of individual implementation strategies or determine whether L2L was effective in helping patients reduce access to firearms for suicide prevention purposes. Measuring the effectiveness of this tool, which was designed to support population-based suicide prevention, would require extending the implementation of L2L to other large health care systems nationwide and conducting other analyses that are designed to measure key functions of suicide prevention practices, including risk identification, engagement in evidence-based risk mitigation and treatment, and supportive care transitions [29]. Finally, L2L is meant to support adult patients at risk of suicide reduce access to firearms and other lethal means; additional tools and strategies are required to support youth at risk of suicide. Notably, there is a similar web-based decision aid that is available for this purpose; "Lock and Protect" was designed to help parents and caregivers reduce access to lethal means for youth suicide risk mitigation [60]. Similarly, the "Safety in Dementia" web-based decision aid was developed to support caregivers in addressing firearm access among individuals with Alzheimer disease and related dementias [61]. Future research should evaluate the implementation of these tools in routine care delivery.

In conclusion, incorporating L2L into the existing system-level safety plan template reduced the effort required to use L2L and was likely the most impactful implementation strategy for increasing clinician adoption and patient reach. However, understanding the value of L2L from the users' perspectives was essential for effectively amplifying the suicide risk mitigation benefits. As rising suicide rates galvanize the urgency of prevention [62], the implementation tools and strategies developed for this project will be useful for health care systems nationwide.

Acknowledgments

This implementation evaluation was funded by Kaiser Permanente's Office of Community Health and Center for Gun Violence Research and Education, as part of its Firearm Injury Prevention Program, and the Centers for Disease Control and Prevention (R01CE003460). The views and opinions expressed in this article are the responsibility of the authors and do not necessarily represent the official views of Kaiser Permanente or the Centers for Disease Control and Prevention. The authors acknowledge that this research would not be possible without the people who receive health care from Kaiser Permanente Washington and all the clinicians and staff who support the organization.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The three dimensions of RE-AIM (Reach, Effectiveness, Adoption, Implementation, Maintenance) selected to inform and evaluate the implementation of "Lock to Live."
[DOCX File, 29 KB - medinform_v12i1e48007_app1.docx]

Multimedia Appendix 2
Patient and clinician interview questions that focused on "Lock to Live" implementation.
[DOCX File, 33 KB - medinform_v12i1e48007_app2.docx]

Multimedia Appendix 3
Elements included in the search for safety plans documented in clinical notes text.
[DOCX File, 31 KB - medinform_v12i1e48007_app3.docx]

Multimedia Appendix 4
Introducing Lock2Live.org: a guide for clinicians.

[DOCX File, 295 KB - medinform_v12i1e48007_app4.docx]

Multimedia Appendix 5
“Lock to Live” Huddlecard.
[DOCX File, 589 KB - medinform_v12i1e48007_app5.docx]

Multimedia Appendix 6
“Lock to Live” video for Kaiser Permanente Washington clinicians.
[MP4 File, 13608 KB - medinform_v12i1e48007_app6.mp4]

References

1. WISQARS — your source for U.S. injury statistics. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/injury/wisqars/facts.html [accessed 2022-04-26]
2. McCracken H, Okuley H, Floyd L. Gun ownership in America. RAND Corporation. URL: https://www.rand.org/research/gun-policy/gun-ownership.html [accessed 2022-01-10]
3. Kalesan B, Villarreal MD, Keyes KM, Galea S. Gun ownership and social gun culture. Inj Prev 2016 Jun;22(3):216-220. [doi: 10.1136/injuryprev-2015-041586] [Medline: 26124073]
4. Parker K, Horowitz JM, Igielnik R, Oliphant JB, Brown A. America’s complex relationship with guns. Pew Research Center. 2017 Jun 22. URL: https://www.pewresearch.org/social-trends/2017/06/22/americas-complex-relationship-with-guns/ [accessed 2022-03-29]
5. Cleveland EC, Azrael D, Simonetti JA, Miller M. Firearm ownership among American veterans: findings from the 2015 National Firearm Survey. Inj Epidemiol 2017 Dec 19;4(1). [doi: 10.1186/s40621-017-0130-y] [Medline: 29256160]
6. Nordstrom DL, Zwerling C, Stromquist AM, Burmeister LF, Merchant JA. Rural population survey of behavioral and demographic risk factors for loaded firearms. Inj Prev 2001 Jun;7(2):112-116. [doi: 10.1136/ip.7.2.112] [Medline: 11428557]
7. Miller M, Zhang W, Azrael D. Firearm purchasing during the COVID-19 pandemic: results from the 2021 National Firearms Survey. Ann Intern Med 2022 Feb;175(2):219-225. [doi: 10.7326/M21-3423] [Medline: 34928699]
8. Anestis MD. Guns and Suicide: An American Epidemic: Oxford University Press; 2018.
9. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Fatal Injury Reports, National, Regional and State, 1981-2017, Web-Based Injury Statistics Query and Reporting System (WISQARS). : Centers for Disease Control and Prevention; 2019.
10. Kellermann AL, Rivara FP, Somes G, et al. Suicide in the home in relation to gun ownership. N Engl J Med 1992 Aug 13;327(7):467-472. [doi: 10.1056/NEJM199208133270705] [Medline: 1308093]
11. Shenassa ED, Rogers ML, Spalding KL, Roberts MB. Safer storage of firearms at home and risk of suicide: a study of protective factors in a nationally representative sample. J Epidemiol Community Health 2004 Oct;58(10):841-848. [doi: 10.1136/jech.2003.017343] [Medline: 15365110]
12. Anglemyer A, Horvath T, Rutherford G. The accessibility of firearms and risk for suicide and homicide victimization among household members: a systematic review and meta-analysis. Ann Intern Med 2014 Jan 21;160(2):101-110. [doi: 10.7326/M13-1301] [Medline: 24592495]
13. Mann JJ, Michel CA. Prevention of firearm suicide in the United States: what works and what is possible. Am J Psychiatry 2016 Oct 1;173(10):969-979. [doi: 10.1176/appi.ajp.2016.16010069] [Medline: 27444796]
14. Ahmedani BK, Simon GE, Stewart C, et al. Health care contacts in the year before suicide death. J Gen Intern Med 2014 Jun;29(6):870-877. [doi: 10.1007/s11606-014-2767-3] [Medline: 24567199]
15. Rowhani-Rahbar A, Simonetti JA, Rivara FP. Effectiveness of interventions to promote safe firearm storage. Epidemiol Rev 2016;38(1):111-124. [doi: 10.1093/epirev/mxv006] [Medline: 26769724]
16. Simonetti JA, Rowhani-Rahbar A, King C, Bennett E, Rivara FP. Evaluation of a community-based safe firearm and ammunition storage intervention. Inj Prev 2018 Jun;24(3):218-223. [doi: 10.1136/injuryprev-2016-042292] [Medline: 28642248]
17. Runyan CW, Becker A, Brandspigel S, Barber C, Trudeau A, Novins D. Lethal means counseling for parents of youth seeking emergency care for suicidality. West J Emerg Med 2016 Jan;17(1):8-14. [doi: 10.5811/westjem.2015.11.28590] [Medline: 26823923]
18. Boggs JM, Beck A, Ritzwoller DP, Battaglia C, Anderson HD, Lindrooth RC. A quasi-experimental analysis of lethal means assessment and risk for subsequent suicide attempts and deaths. J Gen Intern Med 2020 Jun;35(6):1709-1714. [doi: 10.1007/s11606-020-05641-4] [Medline: 32040838]
19. Monuteaux MC, Azrael D, Miller M. Association of increased safe household firearm storage with firearm suicide and unintentional death among US youths. JAMA Pediatr 2019 Jul 1;173(7):657-662. [doi: 10.1001/jamapediatrics.2019.1078] [Medline: 31081861]

20. Boggs JM, Quintana LM, Powers JD, Hochberg S, Beck A. Frequency of clinicians' assessments for access to lethal means in persons at risk for suicide. Arch Suicide Res 2022;26(1):127-136. [doi: 10.1080/13811118.2020.1761917] [Medline: 32379012]
21. Walters H, Kulkarni M, Forman J, Roeder K, Travis J, Valenstein M. Feasibility and acceptability of interventions to delay gun access in VA mental health settings. Gen Hosp Psychiatry 2012;34(6):692-698. [doi: 10.1016/j.genhosppsych.2012.07.012] [Medline: 22959420]
22. Betz ME, Knoepke CE, Siry B, et al. 'Lock to Live': development of a firearm storage decision aid to enhance lethal means counselling and prevent suicide. Inj Prev 2019 Sep;25(Suppl 1):i18-i24. [doi: 10.1136/injuryprev-2018-042944] [Medline: 30317220]
23. Stacey D, Bennett CL, Barry MJ, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2011 Oct 5;10(10):CD001431. [doi: 10.1002/14651858.CD001431.pub3] [Medline: 21975733]
24. Légaré F, O'Connor AC, Graham I, et al. Supporting patients facing difficult health care decisions: use of the Ottawa Decision Support Framework. Can Fam Physician 2006 Apr;52(4):476-477. [Medline: 17327891]
25. Betz ME, Knoepke CE, Simpson S, et al. An interactive web-based lethal means safety decision aid for suicidal adults (Lock to Live): pilot randomized controlled trial. J Med Internet Res 2020 Jan 29;22(1):e16253. [doi: 10.2196/16253] [Medline: 32012056]
26. Boggs JM, Quintana LM, Beck A, et al. "Lock to Live" for firearm and medication safety: feasibility and acceptability of a suicide prevention tool in a learning healthcare system. Front Digit Health 2022 Sep 6;4:974153. [doi: 10.3389/fdgth.2022.974153] [Medline: 36148209]
27. Fast facts. Kaiser Permanente. URL: https://about.kaiserpermanente.org/who-we-are/fast-facts [accessed 2021-11-08]
28. Richards JE, Parrish R, Lee A, Bradley K, Caldeiro R. An integrated care approach to identifying and treating the suicidal person in primary care. Psychiatric Times. URL: https://www.psychiatrictimes.com/view/integrated-care-approach-identifying-and-treating-suicidal-person-primary-care [accessed 2020-01-31]
29. Richards JE, Simon GE, Boggs JM, et al. An implementation evaluation of "Zero Suicide" using normalization process theory to support high-quality care for patients at risk of suicide. Implement Res Pract 2021 Jan 1;2:26334895211011769. [doi: 10.1177/26334895211011769] [Medline: 34447940]
30. Richards JE, Kuo E, Stewart C, et al. Self-reported access to firearms among patients receiving care for mental health and substance use. JAMA Health Forum 2021 Aug 6;2(8):e211973. [doi: 10.1001/jamahealthforum.2021.1973] [Medline: 35977197]
31. National Academies of Sciences, Engineering, and Medicine. Health Systems Interventions to Prevent Firearm Injuries and Death: Proceedings of a Workshop: The National Academies Press; 2019. [doi: 10.17226/25354]
32. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 1999 Sep;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
33. Holtrop JS, Rabin BA, Glasgow RE. Qualitative approaches to use of the RE-AIM framework: rationale and methods. BMC Health Serv Res 2018 Mar 13;18(1):177. [doi: 10.1186/s12913-018-2938-8] [Medline: 29534729]
34. Stetler CB, Legro MW, Wallace CM, et al. The role of formative evaluation in implementation research and the QUERI experience. J Gen Intern Med 2006 Feb;21(Suppl 2):S1-S8. [doi: 10.1111/j.1525-1497.2006.00355.x] [Medline: 16637954]
35. Richards JE, Kuo ES, Whiteside U, et al. Patient and clinician perspectives of a standardized question about firearm access to support suicide prevention: a qualitative study. JAMA Health Forum 2022 Nov 4;3(11):e224252. [doi: 10.1001/jamahealthforum.2022.4252] [Medline: 36416815]
36. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
37. Beebe J. Rapid Qualitative Inquiry: A Field Guide to Team-Based Assessment: Rowman & Littlefield; 2014.
38. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
39. Stanley B, Brown GK. Safety planning intervention: a brief intervention to mitigate suicide risk. Cogn Behav Pract 2012 May;19(2):256-264. [doi: 10.1016/j.cbpra.2011.01.001]
40. Nock MK, Borges G, Bromet EJ, Cha CB, Kessler RC, Lee S. Suicide and suicidal behavior. Epidemiol Rev 2008;30(1):133-154. [doi: 10.1093/epirev/mxn002] [Medline: 18653727]
41. Morgan ER, Gomez A, Rowhani-Rahbar A. Firearm ownership, storage practices, and suicide risk factors in Washington State, 2013-2016. Am J Public Health 2018 Jul;108(7):882-888. [doi: 10.2105/AJPH.2018.304403] [Medline: 29771611]
42. Goldstick JE, Carter PM, Cunningham RM. Current epidemiological trends in firearm mortality in the United States. JAMA Psychiatry 2021 Mar 1;78(3):241-242. [doi: 10.1001/jamapsychiatry.2020.2986] [Medline: 32965479]
43. Ingram DD, Franco SJ. 2013 NCHS Urban-Rural Classification Scheme for Counties. Vital Health Stat 2 2014 Apr(166):1-73. [Medline: 24776070]
44. Ketterer AR, Poland S, Ray K, Abuhasira R, Aldeen AZ. Emergency providers' familiarity with firearms: a national survey. Acad Emerg Med 2020 Mar;27(3):185-194. [doi: 10.1111/acem.13849] [Medline: 31957230]
45. Farcy DA, Doria N, Moreno-Walton L, et al. Emergency physician survey on firearm injury prevention: where can we improve? West J Emerg Med 2021 Feb 8;22(2):257-265. [doi: 10.5811/westjem.2020.11.49283] [Medline: 33856309]

46. Wintemute GJ, Betz ME, Ranney ML. You can: physicians, patients, and firearms. Ann Intern Med 2016 Aug 2;165(3):205-213. [doi: 10.7326/M15-2905] [Medline: 27183181]
47. Richards JE, Hohl SD, Segal CD, et al. "What will happen if I say yes?" perspectives on a standardized firearm access question among adults with depressive symptoms. Psychiatr Serv 2021 Aug 1;72(8):898-904. [doi: 10.1176/appi.ps.202000187] [Medline: 33940947]
48. Khazanov GK, Keddem S, Hoskins K, et al. Stakeholder perceptions of lethal means safety counseling: a qualitative systematic review. Front Psychiatry 2022 Oct 20;13:993415. [doi: 10.3389/fpsyt.2022.993415] [Medline: 36339871]
49. Service O, Hallsworth M, Halpern D, et al. EAST: four simple ways to apply behavioural insights. The Behavioural Insights Team. URL: https://www.bi.team/wp-content/uploads/2015/07/BIT-Publication-EAST_FA_WEB.pdf [accessed 2023-10-12]
50. Nordgren L, Schonthal D. The Human Element: Overcoming the Resistance That Awaits New Ideas: Wiley; 2021.
51. Zero Suicide Institute. Education Development Center. URL: https://solutions.edc.org/solutions/zero-suicide-institute [accessed 2023-10-12]
52. U.S. Department of Health and Human Services (HHS) Office of the Surgeon General and National Action Alliance for Suicide Prevention. 2012 National Strategy for Suicide Prevention: Goals and Objectives for Action: HHS; 2012.
53. Layman DM, Kammer J, Leckman-Westin E, et al. The relationship between suicidal behaviors and Zero Suicide organizational best practices in outpatient mental health clinics. Psychiatr Serv 2021 Oct 1;72(10):1118-1125. [doi: 10.1176/appi.ps.202000525] [Medline: 33730886]
54. Lethal means safety counseling. US Department of Veterans Affairs. URL: https://www.mirecc.va.gov/visn19/lethalmeanssafety/counseling/ [accessed 2023-10-12]
55. TIP 59: improving cultural competence. Substance Abuse and Mental Health Services Administration. 2015. URL: https://store.samhsa.gov/product/TIP-59-Improving-Cultural-Competence/SMA15-4849 [accessed 2023-10-12]
56. Institute of Medicine (US) Committee on Quality of Health Care in America. Crossing the Quality Chasm: A New Health System for the 21st Century: National Academies Press; 2001. [doi: 10.17226/10027]
57. Six domains of healthcare quality. Agency for Healthcare Research and Quality. 2015 Feb. URL: https://www.ahrq.gov/talkingquality/measures/six-domains.html [accessed 2024-03-20]
58. Wosik J, Fudim M, Cameron B, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inform Assoc 2020 Jun 1;27(6):957-962. [doi: 10.1093/jamia/ocaa067] [Medline: 32311034]
59. Simon GE, Stewart CC, Gary MC, Richards JE. Detecting and assessing suicide ideation during the COVID-19 pandemic. Jt Comm J Qual Patient Saf 2021 Jul;47(7):452-457. [doi: 10.1016/j.jcjq.2021.04.002] [Medline: 33994334]
60. Asarnow JR, Zullo L, Ernestus SM, et al. "Lock and Protect": development of a digital decision aid to support lethal means counseling in parents of suicidal youth. Front Psychiatry 2021 Oct 6;12:736236. [doi: 10.3389/fpsyt.2021.736236] [Medline: 34690841]
61. McCarthy V, Portz J, Fischer SM, et al. A web-based decision aid for caregivers of persons with dementia with firearm access (Safe at Home study): protocol for a randomized controlled trial. JMIR Res Protoc 2023 Jan 31;12:e43702. [doi: 10.2196/43702] [Medline: 36719721]
62. Barry E. Following a two-year decline, suicide rates rose again in 2021. The New York Times. 2023 Feb 11. URL: https://www.nytimes.com/2023/02/11/health/suicide-rates-cdc.html#:~:text=Suicide%20increased%20among%20younger%20Black,reported [accessed 2023-02-13]

Abbreviations

C-SSRS: Columbia Suicide Severity Rating Scale
EHR: electronic health record
KPWA: Kaiser Permanente Washington
L2L: Lock to Live
LICSW: licensed clinical social worker
PHQ-9: 9-item Patient Health Questionnaire
RE-AIM: Reach, Effectiveness, Adoption, Implementation, Maintenance

Edited by J Hefner; submitted 07.04.23; peer-reviewed by G Khazanov, J Sung; revised version received 12.10.23; accepted 27.02.24; published 22.04.24.

Please cite as:
Richards JA, Kuo E, Stewart C, Shulman L, Parrish R, Whiteside U, Boggs JM, Simon GE, Rowhani-Rahbar A, Betz ME
Reducing Firearm Access for Suicide Prevention: Implementation Evaluation of the Web-Based "Lock to Live" Decision Aid in Routine Health Care Encounters
JMIR Med Inform 2024;12:e48007
URL: https://medinform.jmir.org/2024/1/e48007
doi:10.2196/48007

Original Paper

Effect of Performance-Based Nonfinancial Incentives on Data Quality in Individual Medical Records of Institutional Births: Quasi-Experimental Study

Biniam Kefiyalew Taye[1,2], MSc; Lemma Derseh Gezie[3], PhD; Asmamaw Atnafu[4], PhD; Shegaw Anagaw Mengiste[5], Prof Dr; Jens Kaasbøll[6], Prof Dr; Monika Knudsen Gullslett[7], Prof Dr; Binyam Tilahun[1], PhD

[1]Department of Health Informatics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

[2]Ministry of Health, The Federal Democratic Republic of Ethiopia, Addis Ababa, Ethiopia

[3]Department of Epidemiology and Biostatistics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

[4]Department of Health System and Policy, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia

[5]Management Information Systems, University of South-Eastern Norway, Drammen, Norway

[6]Department of Informatics, University of Oslo, Oslo, Norway

[7]Faculty of Health & Social Sciences, Science Center Health & Technology, University of South-Eastern Norway, Notodden, Norway

Corresponding Author:
Biniam Kefiyalew Taye, MSc
Ministry of Health, The Federal Democratic Republic of Ethiopia
Zambia street
Addis Ababa
Ethiopia
Phone: 251 910055867
Email: bini.bhi2013@gmail.com

Abstract

Background: Despite the potential of routine health information systems in tackling persistent maternal deaths stemming from poor service quality at health facilities during and around childbirth, research has demonstrated their suboptimal performance, evident from the incomplete and inaccurate data unfit for practical use. There is a consensus that nonfinancial incentives can enhance health care providers' commitment toward achieving the desired health care quality. However, there is limited evidence regarding the effectiveness of nonfinancial incentives in improving the data quality of institutional birth services in Ethiopia.

Objective: This study aimed to evaluate the effect of performance-based nonfinancial incentives on the completeness and consistency of data in the individual medical records of women who availed institutional birth services in northwest Ethiopia.

Methods: We used a quasi-experimental design with a comparator group in the pre-post period, using a sample of 1969 women's medical records. The study was conducted in the "Wegera" and "Tach-armacheho" districts, which served as the intervention and comparator districts, respectively. The intervention comprised a multicomponent nonfinancial incentive, including smartphones, flash disks, power banks, certificates, and scholarships. Personal records of women who gave birth within 6 months before (April to September 2020) and after (February to July 2021) the intervention were included. Three distinct women's birth records were examined: the integrated card, integrated individual folder, and delivery register. The completeness of the data was determined by examining the presence of data elements, whereas the consistency check involved evaluating the agreement of data elements among women's birth records. The average treatment effect on the treated (ATET), with 95% CIs, was computed using a difference-in-differences model.

Results: In the intervention district, data completeness in women's personal records was nearly 4 times higher (ATET 3.8, 95% CI 2.2-5.5; *P*=.02), and consistency was approximately 12 times more likely (ATET 11.6, 95% CI 4.18-19; *P*=.03) than in the comparator district.

Conclusions: This study indicates that performance-based nonfinancial incentives enhance data quality in the personal records of institutional births. Health care planners can adapt these incentives to improve the data quality of comparable medical records, particularly pregnancy-related data within health care facilities. Future research is needed to assess the effectiveness of nonfinancial incentives across diverse contexts to support successful scale-up.

(JMIR Med Inform 2024;12:e54278) doi:10.2196/54278

KEYWORDS

individual medical records; data quality; completeness; consistency; nonfinancial incentives; institutional birth; health care quality; quasi-experimental design; Ethiopia

Introduction

Background

Maternal mortality, a pressing global health concern, is particularly prevalent in low- and middle-income countries [1-5]. The existing research attributes the persistence of maternal deaths largely to inadequate health care quality during labor, delivery, and immediate postpartum care in health facilities [6,7]. Almost every low- and middle-income country implements the Routine Health Information System (RHIS) to address this challenge [8-10]. The RHIS has gained prominence for its practical roles in improving the quality of services, including (1) facilitating evidence-based action by enabling the early detection of pregnancy-related complications, (2) serving as a repository for clients' data to ensure the continuity of pregnancy-related care, and (3) functioning as the primary data source essential for health monitoring and evaluation at all levels of the public health system [11-16]. Despite its potential, the performance of RHIS remains suboptimal, primarily because of incomplete and inaccurate data, hindering its effective use by decision makers [17-20].

In Ethiopia, the introduction of the RHIS dates back to 2008 [21,22]. Ongoing efforts are in place to enhance the data quality of the RHIS in Ethiopia through interventions such as the Performance Monitoring Team (PMT), lot quality assurance sampling (LQAS), and the Capacity Building and Mentorship Program (CBMP) [23-25]. However, despite these efforts, the quality of RHIS data still lags in Ethiopia [15,26]. This challenge is pertinent to institutional birth, as shown by some previous studies in Ethiopia. For instance, a study [27] reported a completeness rate of only 18.4%. Another study from Ethiopia found that 66% of health facilities managed to produce accurate data within an acceptable range [28]. Furthermore, comparing the data from health facilities with external sources such as the Ethiopian Demographic Health Survey reveals concerns regarding data quality in Ethiopian RHIS [26].

Incentives and Its Impact on Health Care Quality

Previous studies have shown that offering incentives for personnel responsible for data collection and management can improve data quality in the RHIS. According to some studies, incentives are essential for addressing the negative attitudes and values that undermine data quality within the RHIS, which are primary challenges to achieving desired quality of RHIS data [29].

The effectiveness of incentives in health care is grounded in theoretical and empirical evidence. Theories like the theory of planned behavior emphasize the connection between motivation and improved health care quality [30]. Some studies demonstrated that incentive-based interventions can predict up to 48% of desired health care behavior [31,32].

Despite the growing interest in using incentives in health care [32-38], determining the most effective approach remains a research priority. Incentives can be financial [39,40] or nonfinancial. Financial incentives have been extensively studied globally [33,37-39,41-43], but there is limited evidence supporting their consistent impact on health care quality [40]. Some studies have even cited the counterproductive effects of financial incentives on health care [44].

Compared to financial incentives, the impact of nonfinancial incentives on health care quality has been minimally studied [45]. Nonfinancial incentive schemes offer noncash rewards or benefits to motivate recipients using approaches that involve recognition through public profiling or reporting; career advancement opportunities; providing certificates to top performers; and ensuring improved working conditions, such as vacations, grading systems, and packaging interventions with in-kind items [46-53]. Previous studies have demonstrated the effectiveness of nonfinancial incentives in enhancing the quality of health services. For instance, nonfinancial incentive schemes in the United States, India, El Salvador, and Tanzania have been reported to enhance the performance of health care providers, including enhanced root cause data analysis of medical errors, expanded community outreach services, better maternal and child care services, and higher quality health care consultations [50,51,54,55].

Objectives

This study aimed to evaluate the effect of performance-based nonfinancial incentives (PBNI) on enhancing the quality of institutional birth data, measured by the completeness and consistency of data within women's individual medical records (IMRs).

Methods

Study Design and Period

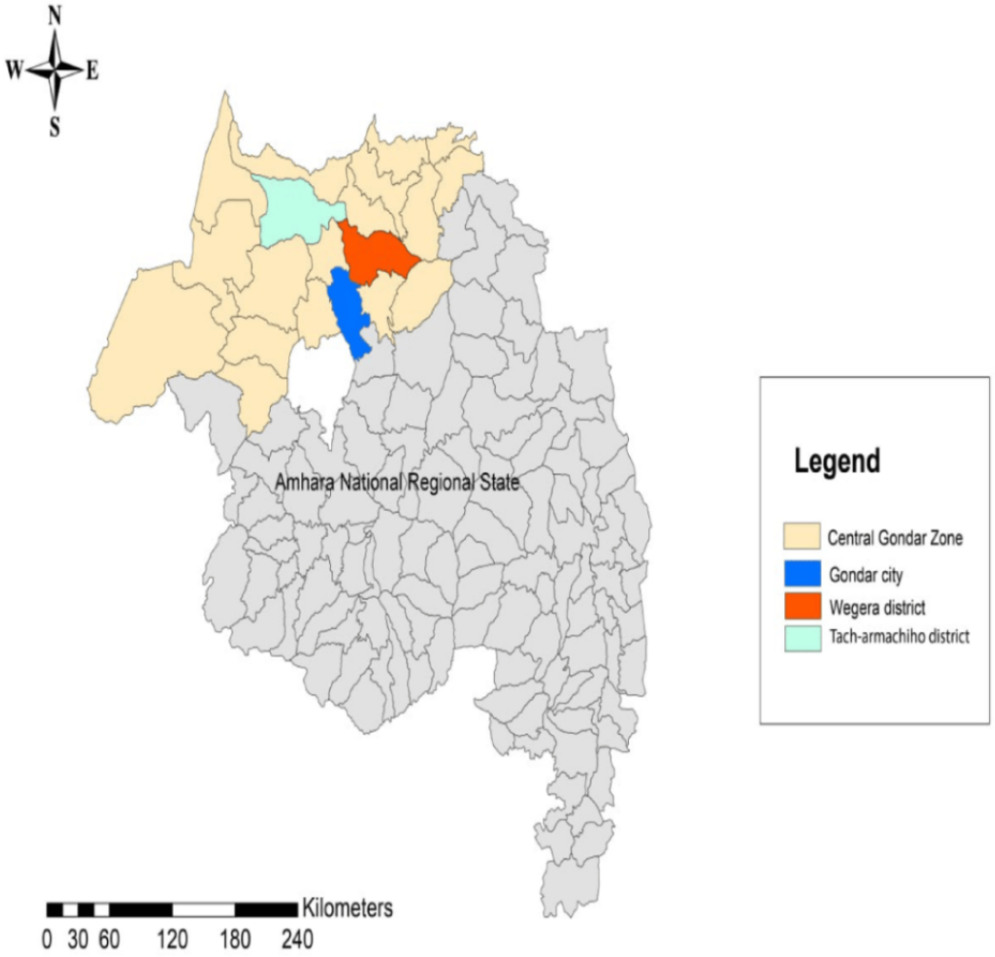
This study used a quasi-experimental design with a comparator group in the pre-post period to examine the effect of PBNI on the data quality of IMRs of institutional births. A cross-sectional survey within an institutional setting was used to review institutional birth–related medical records. The study included the IMRs of women who gave birth within 6 months before (April to September 2020) and after (February to July 2021) the intervention.

Study Setting

This study was conducted in the Amhara National Regional State, specifically in the Central Gondar Zone districts of "Wegera" and "Tach-armachiho" in northwest Ethiopia (Figure 1). According to the Ethiopian Central Statistical Agency projection, the total population in both districts was approximately 398,350 in 2021, with an estimated 93,214 women in the reproductive age group (15-49 years) [56,57].

The districts hosted 2 primary hospitals, 13 health centers, and 72 health posts. The intervention district had 678 health care providers and 215 support staff, whereas the comparator district had 202 health care providers and 141 support staff [58]. These districts were chosen for their involvement in the CBMP, a collaborative initiative between the Ethiopian Ministry of Health and the University of Gondar [24,59,60]. As part of the CBMP, the University of Gondar provides ongoing technical support, including training, supervision, and mentorship to health facilities in both districts, to enhance data quality and information use [59]. This study evaluated the effect of PBNI in the "Wegera" district, with the "Tach-armachiho" district serving as a comparator.

Figure 1. Map of the study area.



Intervention

Intervention Aim and Design

The PBNI intervention was implemented in the "Wegera" district between October 2020 and July 2021 to improve the data quality and use of the RHIS. PBNI was designed as a package with multiple nonfinancial incentive components. Incentives were offered across 3 levels: health facilities, departments, and individual health care workers. Individual health workers were offered nonfinancial incentives, including smartphones, flash disks, power banks, and scholarships. Desktop computers were offered at the department and health facility levels. Health workers, departments, and health facilities that earned nonfinancial incentives in each round were awarded certificates of recognition [61].

Target Areas for PBNI

The study included 6 (75%) out of the 8 health centers in "Wegera." The departments involved in the PBNI program include Maternal and Child Health, Outpatient Department, under-5 Outpatient Department, the Health Management Information System (HMIS), and the Medical Record Unit (MRU). Health workers who participated in the PBNI program included various experts, such as medical record personnel, health IT (HIT) personnel, health officers, midwives, nurses, and personnel involved in laboratories and pharmacies.

Awardees Selection Procedures

The selection of the best performers was conducted through 2 approaches: a subjective and an objective approach.

Subjective Approach

The subjective approach involved requesting management authorities in the intervention district to nominate the best-performing employees. The subjective approach was chosen owing to practical constraints, as the quantitative measurement of all health workers' performance was infeasible owing to limited resources. Accordingly, the number of potential

awardees was reduced to a manageable level, allowing us to concentrate on objectively evaluating the candidates.

The subjective approach was conducted in 2 phases. In the initial phase, health office department managers in the intervention district nominated 12 individuals, selecting 2 from each of the 6 participating health centers. The second phase mirrored the first phase, except that the selection process took place at the level of each health center, where the heads of each health center were tasked with nominating the best performers. With 2 nominees selected by the heads of each health center, another 12 individuals were identified. Consequently, 24 individuals were identified using a subjective approach.

Objective Approach

Previous research indicates that effective health care incentives depend on rewarding specific performance [62,63]. In this study, for the purpose of incentivizing 3 entities—health centers, departments, and individual health workers—we used a flexible approach that used objective measures to identify the best performers. For health centers, 14 quantitative performance measures, each of which was established with specific targets and points to be earned, were used (Textbox 1). The allocation of point values and performance targets took priority for the RHIS activities, as defined by the Ethiopian Ministry of Health [22].

The performance of departments and the 24 individuals selected during the subjective phases was objectively evaluated, aligning the 14 quantitative performance measures with their relevant roles and job descriptions. Further details on the performance measures used are described in prior studies [58,61,64,65].

Textbox 1. The performance indicators used to determine the awardees of nonfinancial incentives, northwest Ethiopia, 2021.

Indicators and points (total points: 90)

1. Source documents completeness rate: 10
2. Report timeliness: 5
3. Lot quality assurance sampling performance: 6
4. Data consistency among registers and reports: 12
5. Health centers established by Performance Monitoring Team: 8
6. Action plan implemented regularly: 10
7. Conducted internal supervision: 5
8. Gaps identified and prioritized by Performance Monitoring Team: 5
9. Conducted root cause analysis: 5
10. Feedback provided for case teams by health centers: 4
11. Number of feedback entries provided to health posts by health centers 5:
12. Information display status: 5
13. Report completeness: 5
14. Consistency among medical records: 5

The Awarding Processes

Initially, the team from the University of Gondar visited the health office department and health centers in the intervention district to communicate the commencement of the program. During this announcement phase, a banner illustrating the nonfinancial rewards was displayed within the compounds of the health facilities. Nonfinancial incentives were offered to the recipients through 3 ceremonial award programs that took place bimonthly. The attendees of the PBNI ceremonial award include representatives from the University of Gondar, Federal Ministry of Health, Amhara Regional Health Bureau, Central Gondar Zone, and “Wegera” District Health Office departments. These representatives comprised health experts and administrative personnel. Officials from the Federal Ministry of Health and University of Gondar rewarded the top-3 individuals, departments, and health centers. Certificates of recognition were presented to awardees during these bimonthly forums. Ceremonial events were also accompanied by presentations detailing the performance measurement and award selection procedures by the professionals from the University of Gondar.

Study Participants

Overview

The participants in this study were women who had given birth in the health centers at the study sites. The IMR sets of these individuals were examined. Thus, for each woman, there would be a set of 3 types of records: delivery register, integrated individual folder (IIF), and integrated card. These 3 sets of IMRs were combined to form a single study cohort. The 3 types of IMRs evaluated in this study, designed to record institutional birth data following the guidelines established in Ethiopia [66], are described in the following sections.

Integrated Card

The integrated card captures data on pregnant women throughout antenatal, labor, delivery, and postnatal care. It facilitates the recording of medical histories, physical examination results,

and other clinical data for both women and newborns, allowing health care providers to complete it upon birth.

Delivery Register

The delivery register is a serial-long register designed to contain a list of all women who give birth at the facility, with data abstracted from the integrated card.

Women's IIF

The IIF is designed to consolidate the entirety of a woman's personal records, including the integrated card. It ensures the convenient access to comprehensive medical data, with the front section containing personal identification data filled out during registration and the inner part featuring a summary sheet completed by service providers after each visit.

Sample Size Calculation

The required sample size for this study was calculated using StatCalc (Epi Info version 7.0; Centers for Disease Control and Prevention), incorporating assumptions to detect differences in completeness and consistency rates between the intervention and comparator groups. The assumptions included 80.3% completeness and 29.5% consistency from a prior unpublished pilot study (Taye, BK, unpublished data, September 2021), a 1:1 ratio of the intervention to the comparator group, a 5% anticipated change in the intervention group [55], 80% power, 95% CI, and a 10% nonresponse rate. Separate calculations yielded approximately 1969 participants (985 in each group) for completeness and approximately 3007 (1504 in each group) for consistency. From the 2 computed samples, we chose 1969 participants, considering the available resources for feasibility.

Inclusion and Exclusion Criteria

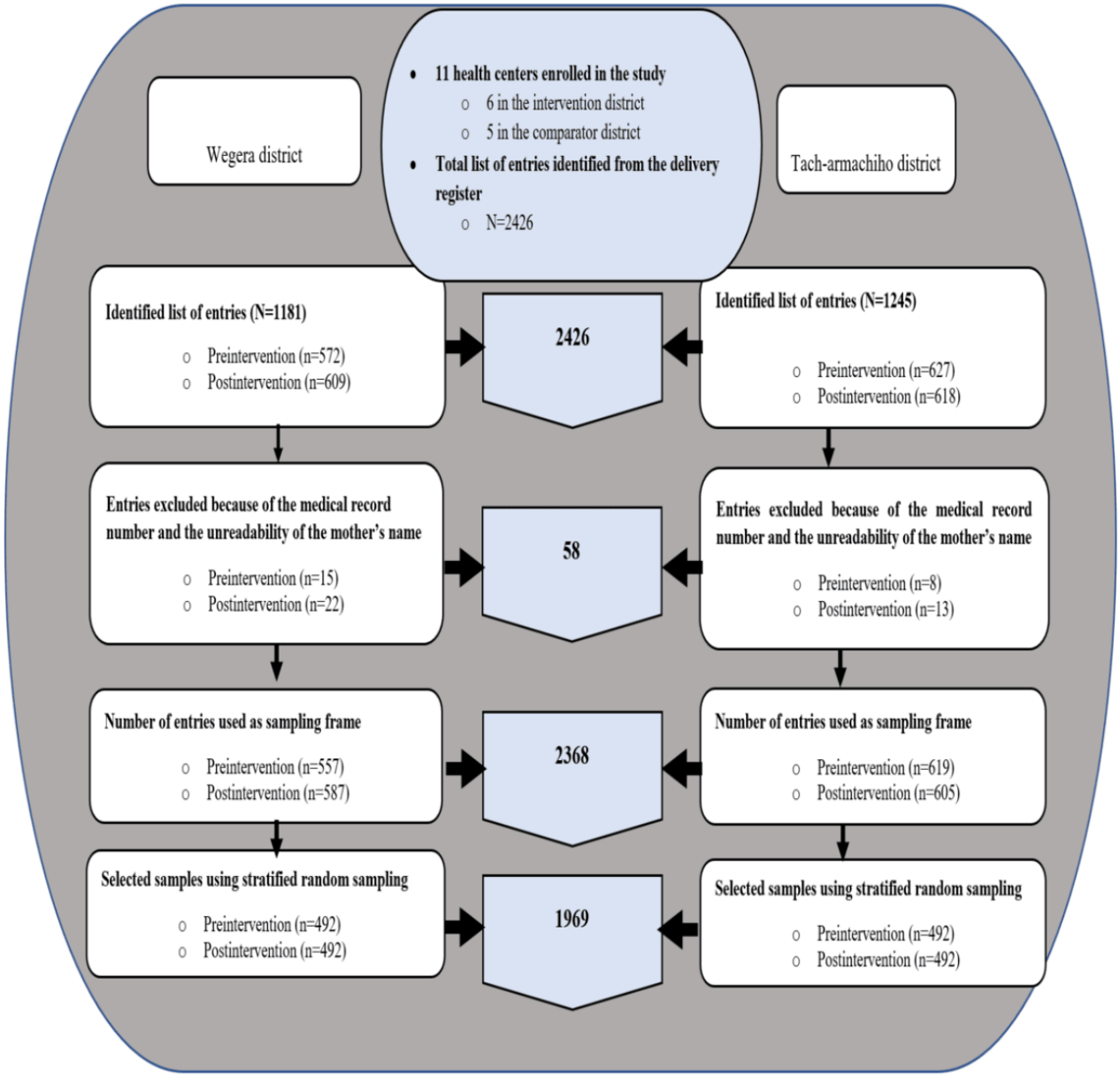
A list of women who gave birth during the study period was used as a sampling frame to select the study sample. Three distinct IMRs in combination (delivery register, IIF, and integrated card) were necessary for each participant. The inclusion criteria in this study mandated that women's "medical record numbers" (MRNs) and the mother's full name be legibly recorded on the delivery register. The readability of these 2 variables was necessary to match and retrieve women's IIF data from the MRU.

Sampling Procedure

A total of 11 health centers were included in the pre- and postintervention periods, of which 6 (55%) were in the intervention district and 5 (45%) were in the comparator district. In total, 2426 women's entries were identified from the delivery register across the baseline and end line periods.

Of the 2426 entries, we excluded 58 (2%) due to unreadability of the MRNs and the full names of the mothers, resulting in 2368 (98%) entries of women on the delivery register. The list of 2368 women in the delivery register was used as a sampling frame to select 1969 (83.1%) samples through stratified random sampling. Subsequently, the women's IIF was retrieved from the MRU (Figure 2).

Figure 2. A diagram illustrating the process of study sample selection.



Variables and Measurements

Outcome Variables

Completeness

Completeness refers to the presence of data elements in the targeted data set [67]. Within the RHIS, assessing the completeness of data elements mandates the presence of its corresponding services or medical procedures in health facilities [68]. Accordingly, this research evaluated the completeness of the 33 data elements where the guideline necessitates their recording [66]. Of the 33 data elements, 16 (48%) were from the integrated card (mother's name, gravida, para, MRN, date and time of admission, ruptured membranes, date and time of delivery, mode of delivery, placenta details, newborn's sex, newborn outcome, single or multiple births, term or preterm status, and the name and signature of providers), 7 (21%) were from the individual folder (facility name, registration date, client's sex, mother's age, date of delivery, Department-provided service, and serial number), and 10 (30%) were from the delivery register (serial number, age, address, date and time of delivery, newborn outcome, mode of delivery, maternal status, newborn's sex, and the name and signature of providers).

Consistency

According to previous research [69], consistency is a measure of data accuracy (the extent to which data elements accurately represent the true numbers), commonly assessed in RHIS through data verification (agreement of data among data sources). This research crosschecked the agreement between the data elements in the source documents to assess consistency. The delivery register served as the gold standard, allowing a comparison with 12 data elements abstracted from the IIF and integrated card, including the serial number, MRN, mother's name, delivery date and time, mode of delivery, newborn's sex, provider's name and signature, Apgar score, newborn's weight, mother's HIV test acceptance, and mother's HIV test results.

In this study, data collectors judged the completeness and consistency of data elements, and the decision adhered to the guidelines [66]. For completeness, data collectors observed the presence of data elements in personal records and declared and coded each data element as 1 for "Yes" if recorded and 0 for "No" if unrecorded. Regarding consistency, data collectors verified elements from comparable records for agreement, coding them as 1 for "Yes" if consistently recorded and 0 for "No" otherwise.

Independent Variables

The characteristics of health facilities supposed to be associated with the data quality of institutional birth were included in the analysis, drawing from pertinent literature and guidelines in the field [22,66,70-72]. These independent variables were the presence of HIT personnel, availability of data recording tools, availability of trained providers, supportive supervision from higher officials, existence of PMT, PMT per membership standard, monthly PMT meetings, monthly conducted LQAS, conducted root cause analysis (RCA) on the identified gaps, internal supervision, and availability of HMIS guidelines. Details of the measurement and data management for outcomes and independent variables have been outlined in a previous study [73].

Data Collection

In total, 11 data collectors, HIT personnel, and related health sciences graduates were recruited for the data collection. The data collectors received a 3-day training that covered the study's objectives, methodology, and ethical considerations. The principal investigator and the other 2 supervisors supervised the data collection process.

Statistical Analysis

Data were entered into the EpiData (version 3.1; Epi Data Association) and exported into Stata (version17.0; StataCorp) for analysis.

Descriptive Statistics

The frequency distribution of women's IMRs is presented based on the characteristics of the health facilities. Pearson chi-square tests were computed to assess the comparability of the data on the study participants' baseline characteristics between the intervention and comparator groups.

The completeness and consistency proportions were computed for specific data elements, with the average rates calculated for the study participants. The completeness proportion of data elements was determined by dividing the number of participants with recorded data elements by the total number of study participants. Meanwhile, the mean completeness proportion was calculated by dividing the number of data elements recorded per participant by the total expected number of data elements. Likewise, the consistency proportion of data elements was determined by dividing the number of participants with consistently recorded data elements by the total number of study participants, and the mean consistency proportion was calculated by dividing the number of data elements consistently recorded per participant by the total expected number of data elements.

Changes in completeness and consistency proportions between the study groups were compared by computing absolute differences and their corresponding 95% CIs, using a 2-sample proportion test.

Difference-in-Differences Analysis

This study used difference-in-differences (DID) to estimate the average effect of PBNI on data completeness and consistency [74,75]. The average treatment effect on the treated (ATET) was computed using the DID model. The DID model used can be described as follows:

(1)

where γ_i represents the outcome (either completeness or consistency) for each study participant. The variable 1(g = intervention) is an indicator for the study group, taking the value of "1" if the participant belongs to the intervention group and "0" otherwise. Similarly, the variable 1(t = post) is a binary indicator, taking the value of "1" for the participants sampled during the end line period and 0 otherwise. $Z_{igt}\beta$ represents the independent variables, that is, the characteristics of health facilities included in the DID model. D represents the intervention in this study (PBNI). σ represents the coefficient of average treatment effect on the intervention group (ATET), providing estimates of the average effect of the PBNI on the outcome variables, and  are the residual errors.

Ethical Considerations

Ethics clearance was obtained from the University of Gondar's Institutional Ethical Review Board (RNO: V/P/RCS/05/861-2021). As this study used medical records rather than human participants, obtaining informed consent from the participants was not feasible.

Results

Description of Women's IMRs by Health Facilities' Baseline Characteristics

In this study, 91.92% (1810/1969) of the samples met the analysis criteria for the coexistence of all 3 distinct IMRs. Of the analysis sample, 49.78% (901/1810) were sampled from the intervention district and 50.22% (909/1810) from the comparator district, whereas 49.56% (897/1810) were enrolled at baseline and 50.44% (913/1810) were enrolled at end line.

Among the baseline samples involved in the analysis, 51% (458/897) were enrolled from the intervention district and 48.9% (439/897) were from the comparator district. When comparing study participants according to baseline health facility characteristics, the distribution was identical (897/897, 100%) across the following variables: the presence of HIT personnel, the existence of PMT, the PMT established per membership standard, monthly PMT meetings in the last 6 months, and monthly LQAS in the previous 6 months.

Most of the IMRs (425/458, 92.8%) in the intervention district were from health centers where the data recording tools were fully available, compared with 12.3% (54/439) in the comparator

district (*P*<.001). In the intervention district, 69% (316/458) of the IMRs were from health centers where most providers had received training on data quality, compared with 65.6% (288/439; *P*=.28) in the comparator district. In the intervention district, 39.7% (182/458) of the IMRs were from health centers with at least 3 supportive supervision visits by higher officials, compared with 65.6% (289/439; *P*<.001) in the comparator district. Nearly one-third of the IMRs (145/458, 31.6%) in the intervention district were from health centers that conducted at least 2 internal supervisions, compared with 4.8% (21/439; *P*<.001) in the comparator district. Less than three-fourth of the IMRs (303/458, 66.2%) in the intervention district were from health facilities that conducted RCA at least 3 times, compared with 25.2% (111/439; *P*<.001) in the comparator district. In the intervention district, 92.8% (425/458) of the IMRs were from health centers with fully available HMIS guidelines, compared with 91.8% (403/439; *P*=.58) in the comparator district (Table 1).

Table 1. Baseline comparison between the study groups by the characteristics of health facilities, northwest Ethiopia, 2021.

Variables	Intervention (N=458), n (%)	Comparator (N=439), n (%)	*P* value
Presence of health information technology personnel	458 (100)	439 (100)	<.001
Existence of PMT[a]	458 (100)	439 (100)	<.001
The PMT per membership standard	458 (100)	439 (100)	<.001
Monthly PMT meeting	458 (100)	439 (100)	<.001
Monthly conducted lot quality assurance sampling	458 (100)	439 (100)	<.001
Availability of data recording tools			
Fully	425 (92.8)	54 (12.3)	<.001
Partially	33 (7.2)	385 (87.7)	<.001
Availability of trained providers			
Mostly	316 (69)	288 (65.6)	.28
Partially	142 (31)	151 (34.4)	.28
Supportive supervisions from higher officials			
<3 times	276 (60.2)	151 (34.4)	<.001
At least 3 times	182 (39.7)	289 (65.6)	<.001
Conducted root cause analysis on the identified gap			
Yes	303 (66.2)	111 (25.2)	<.001
No	155 (33.8)	328 (74.7)	<.001
Internal supervision			
At least 2 times	145 (31.6)	21 (4.8)	<.001
<2 times	313 (68.3)	418 (95.2)	<.001
Availability of the Health Management Information System guidelines			
Fully available	425 (92.8)	403 (91.8)	.58
Partially available	33 (7.2)	36 (8.2)	.58

[a]PMT: Performance Monitoring Team.

Specific Data Elements Completeness Across Study Groups

Table 2 compares the intervention and comparator districts regarding the completeness of specific data elements across the 3 IMRs. Concerning the data elements from the integrated card, the "Name of the mother" showed 95.6% (861/901) completeness in the intervention district compared with 92.6% (842/909; *P*=.004) in the comparator district. The completeness proportion of "Gravida" was 91.7% (826/901) in the intervention district, compared with 90.4% (822/909; *P*=.18) in the comparator district. The completeness proportion of "MRN" was 92.7% (835/901) in the intervention district, compared with 84.9% (771/909; *P*<.001) in the comparator district. The completeness proportion of "Time of Admission" was 92.5% (833/901) in the intervention district, compared with 88.7% (806/909; *P*=.003) in the comparator district. The completeness proportion of "time of delivery" was 94.9% (855/901) in the intervention district, compared with 89.6% (814/909; *P*<.001) in the comparator district. The completeness proportion of "Name and signature of providers" was 90.1% (812/901) in the intervention district, compared with 86.8% (789/909; *P*=.01) in the comparator district.

In the IIF, the "date of delivery" completeness proportion was 39.3% (354/901) in the intervention district, compared with 29.4% (268/909; *P*<.001) in the comparator district. The completeness proportion of "Date of registration" was 99.8%

(900/901) in the intervention district, compared with 93.4% (849/909; *P*<.001) in the comparator district. The completeness proportion of "Department-provided service" was 37.1% (334/901) in the intervention district, compared with 29.4% (267/909; *P*<.001) in the comparator district. The completeness proportion of "Serial Number" was 35.2% (318/901) in the intervention district, compared with 31% (282/909; *P*=.03) in the comparator district.

Regarding the data elements in the delivery register, the "Serial Number" and "Age" were found to be recorded for all study participants across the intervention and comparator districts. In the intervention district, the "time of delivery" showed 92.7% (835/901) completeness, compared with the comparator district, which showed 62.6% (571/909; *P*<.001) completeness. In the intervention district, the "date of delivery" showed 99.7% (899/901) completeness, compared with the comparator district, which showed 95.3% (867/909; *P*<.001) completeness. The completeness proportion of "sex of newborn" was 99.4% (896/901) in the intervention district, compared with 99% (900/909; *P*=.14) in the comparator district. The "Name and signature of providers" completeness proportion was 78.3% (706/901) in the intervention district, compared with 80.9% (736/909; *P*=.92) in the comparator district.

Table 2. Specific data elements' completeness in individual medical records of institutional births across intervention and comparator districts, northwest Ethiopia, 2021.

Completeness	Intervention (N=901), n (%)	Comparator (N=909), n (%)	Difference[a] (95% CI)	*P* value[b]
Integrated card				
Name of the mother	861 (95.6)	842 (92.6)	2.9 (0.76 to 5)	.004
Gravida	826 (91.7)	822 (90.4)	1.2 (−1.48 to 3.8)	.18
Para[c]	858 (95.2)	820 (90.2)	5 (2.6 to 7.49)	<.001
Medical record number	835 (92.7)	772 (84.9)	7.7 (4.8 to 10.6)	<.001
Date of admission	841 (93.3)	822 (90.4)	2.9 (0.4 to 5.4)	.01
Time of admission	833 (92.5)	806 (88.7)	3.8 (1 to 6.4)	.003
Ruptured membranes	664 (73.7)	760 (83.6)	9.9 (−13.6 to −6.26)	>.99
Date of delivery	861 (95.6)	843 (92.7)	2.8 (0.6 to 4.9)	.005
Time of delivery	855 (94.9)	814 (89.6)	5.3 (2.8 to 7.89)	<.001
Mode of delivery	834 (92.6)	806 (88.7)	3.8 (1.2 to 6.6)	.002
Placenta	842 (93.5)	804 (88.5)	5 (2.4 to 7.6)	<.001
Sex of the newborn	851 (94.5)	813 (89.4)	5 (2.5 to 7.5)	<.001
Newborn outcome	850 (94.3)	822 (90.4)	3.9 (1.5 to 6.3)	<.001
Single or multiple	756 (83.9)	802 (88.2)	4.3 (−7.5 to −1.1)	>.99
Term or preterm	786 (87.2)	801 (88.1)	0.8 (−3.9 to 2.1)	.72
Name and signature	812 (90.1)	789 (86.8)	3.3 (0.38 to 6.26)	.01
Integrated individual folder				
Name of the facility	887 (98.4)	807 (88.7)	9.6 (7.4 to 11.8)	<.001
Date of registration	900 (99.8)	849 (93.3)	6.4 (4.8 to 8.1)	<.001
Sex of the client	899 (99.7)	858 (94.3)	5.4 (3.8 to 6.9)	<.001
Age of the mother	843 (93.5)	860 (94.6)	1 (−3.2 to 1.1)	.83
Date of delivery	354 (39.2)	268 (29.4)	9.8 (5.4 to 14.2)	<.001
Department-provided service	334 (37)	267 (29.3)	7.6 (3.4 to 12)	<.001
Serial number	318 (35.2)	282 (31)	4.3 (−0.01 to 8.6)	.03
Delivery register				
Serial number	901 (100)	909 (100)	—[d]	<.001
Age	901 (100)	909 (100)	—	<.001
Address	900 (99.8)	900 (99)	0.8 (0.2 to 1.6)	.005
Date of delivery	899 (99.7)	867 (95.3)	4.4 (2.9 to 5.8)	<.001
Time of delivery	835 (92.6)	571 (62.8)	29.8 (26.2 to 33.4)	<.001
Newborn outcome	897 (99.5)	905 (99.5)	—	<.001
Mode of delivery	898 (99.6)	899 (98.8)	0.8 (−0.01 to 1.5)	.05
Maternal status	897 (99.6)	909 (100)	0.4 (−0.8 to –0.01)	.98
Sex of the newborn	896 (99.4)	900 (99)	0.4 (−0.37 to 1.24)	.14
Name and signature	706 (78.3)	736 (80.9)	2.6 (−6.31 to 1.09)	.92

[a]The absolute difference is calculated by subtracting the completeness proportion of the comparator group from that of the intervention group.

[b]*P* value based on 2 independent sample proportion tests.

[c]A number of times a woman has given birth to a viable child.

[d]No difference among intervention and comparator group.

Average Data Completeness Across the Pre- and Postintervention Periods

For the integrated card, the average completeness increased from 86.2% (95% CI 83.9%-88.57%) at the baseline to 96.6% (95% CI 96%-97.1%) at the end line in the intervention district; however, in the comparator district, it showed a decrease from 91.1% (95% CI 89.4%-92.7%) at the baseline to 87% (95% CI 84.2%-89.7%) at the end line.

The average completeness of the IIF increased from 58.9% (95% CI 57.6%-60.2%) at the baseline to 85.3% (95% CI 83.6%-87%) at the end line in the intervention district, whereas the comparator district showed a change from 63.5% (95% CI 61.5%-65.4%) to 68.1% (95% CI 65.6%-70.6%).

In the intervention district, the mean completeness proportion of the delivery register increased from 94.6% (95% CI 93.9%-95.2%) at the baseline to 99.3% (95% CI 99%-99.5%) at the end line. In comparison, the comparator district showed a change from 93.5% (95% CI 92.9%-94%) to 93.6% (95% CI 92.9%-94.3%).

In the intervention district, the average data completeness proportion across the 3 individual IMRs was 82.9% (95% CI 81.88%-4.1%) at the baseline, and it increased to 95% (95% CI 94.6%-95.5%) at the end line. In the comparator district, the average data completeness proportion across the 3 IMRs was 86% (95% CI 84.96%-86.97%) at the baseline but decreased to 84.97% (95% CI 83.28%-86.66%) at the end line (Multimedia Appendix 1).

Effect of PBNI on Data Completeness

In the intervention district, the "integrated card" resulted in an average 2.6 percentage-point increase in completeness compared with the comparator district (ATET 2.67, 95% CI 0.7-4.4; *P*=.04). On average, the intervention district showed a 3.8 percentage-point increase in the completeness of the delivery register compared with the comparator district (ATET 3.8, 95% CI 2.9-4.8; *P*=.01). The intervention district showed a 6.8 percentage-point increase in the average completeness of the IIFs compared with the comparator district (ATET 6.8, 95% CI 4.55-9; *P*=.02). Overall, on average, the intervention district showed a 3.8 times higher chance of complete recording of IMRs compared with the comparator district (ATET 3.8, 95% CI 2.2-5.5; *P*=.02; Table 3).

Table 3. Effect of performance-based nonfinancial incentives on the data completeness in individual medical records of institutional births, northwest Ethiopia, 2021.

Completeness	Intervention, mean (SD)	Comparator, mean (SD)	Intervention effect, ATET[a] (95% CI)[b]	*P* value
Integrated card	91.3 (18.8)	88.9 (24.9)	2.6 (0.7-4.4)	.04
Integrated individual folder	71.9 (21.1)	65.8 (24.8)	6.8 (4.6-9)	.02
Delivery register	96.8 (5.8)	93.6 (6.9)	3.8 (2.9-4.8)	.01
Overall	88.8 (11.2)	85.4 (15.3)	3.8 (2.2-5.5)	.02

[a]ATET: average treatment effect on the treated.

[b]ATET estimates adjusted for covariates (presence of health information technology personnel, availability of data recording tools, availability of trained providers, supportive supervision from higher officials, existence of the Performance Monitoring Team [PMT], PMT per membership standard, monthly PMT meeting, monthly conducted lot quality assurance sampling, conducted root cause analysis, internal supervision, and availability of Health Management Information System guidelines).

Consistency of Specific Data Elements Across Study Groups

Regarding the delivery register and IIF, the "date of delivery" showed a consistency proportion of 82.3% (742/901) in the intervention district, compared with 58.7% (534/909; *P*<.001) in the comparator district. The "Serial Number" showed a consistency proportion of 42.1% (380/901) in the intervention district, compared with 45.5% (414/909; *P*=.92) in the comparator district.

Comparing the delivery register and integrated card, the "MRN" exhibited a consistency proportion of 87% (784/901) in the intervention district, compared with 74.9% (681/909; *P*<.001) in the comparator district. The "time of delivery" showed a consistency proportion of 88.2% (795/901) in the intervention district, compared with 56.7% (516/909; *P*<.001) in the comparator district. The "Name and signature of providers" showed a consistency proportion of 89.5% (807/901) in the intervention district, compared with 77.7% (707/909; *P*<.001) in the comparator district. The "newborn weight" had a consistency proportion of 85.7% (773/901) in the intervention district, compared with 82.1% (746/909; *P*=.01) in the comparator district. The "Women's HIV test accepted" showed a consistency proportion of 39.9% (360/901) in the intervention district and 40.5% (368/909; *P*=.59) in the comparator district (Table 4).

Table 4. Consistency of specific data elements across the intervention and comparator districts, northwest Ethiopia, 2021.

Consistency	Intervention (N=901), n (%)	Comparator (N=909), n (%)	Difference[a] (95% CI)	*P* value[b]
Delivery register vs integrated individual folder				
Date of delivery	742 (82.3)	534 (58.7)	23.6 (19.55 to 27.66)	<.001
Serial number	380 (42.1)	414 (45.5)	3.36 (−7.93 to 1.2)	.92
Delivery register vs integrated card				
Medical record number	784 (87)	681 (74.9)	12.09 (8.52 to 15.66)	<.001
Name of the mother	828 (91.8)	709 (77.9)	13.90 (10.67 to 17.12)	<.001
Date of delivery	859 (95.3)	803 (88.3)	6.99 (4.50 to 9.49)	<.001
Time of delivery	795 (88.2)	516 (56.7)	31.46 (27.62 to 35.31)	<.001
Mode of delivery	831 (92.2)	796 (87.5)	4.66 (1.89 to 7.42)	<.001
Sex of the newborn	846 (93.9)	806 (88.6)	5.22 (2.64 to 7.81)	<.001
Name and signature	807 (89.5)	707 (77.7)	11.78 (8.42 to 15.14)	<.001
Apgar score	781 (86.6)	746 (82.1)	4.61 (1.27 to 7.95)	.003
Newborn weight	773 (85.7)	746 (82.1)	3.72 (0.34 to 7.1)	.01
Women’s HIV test accepted	360 (39.9)	368 (40.4)	0.52 (−5.04 to 3.98)	.59
Women’s HIV test result	615 (68.2)	381 (41.9)	26.34 (21.92 to 30.76)	<.001

[a]The absolute difference is calculated by subtracting the consistency proportion of the comparator group from that of the intervention group.

[b]*P* value based on 2 independent sample proportion tests.

Pre- and Postintervention Changes in Average Data Consistency

In the intervention district, the average consistency proportion increased from 71.6% (95% CI 69.6%-73.6%) to 89.2% (95% CI 88.2%-90.2%) after the intervention. In the comparator district, it increased from 68% (95% CI 66.2%-69.8%) to 70.8% (95% CI 67.9%-73.6%) post intervention. Overall, the average consistency proportion increased from 69.8% (95% CI 68.5%-71.2%) to 79.6% (95% CI 78%-81.3%) after the intervention.

Effect of PBNI on Data Consistency

On average, the intervention district showed an 11.2 percentage-point increase in the consistency of data among the delivery register and IIF compared with the comparator district (ATET 11.2; 95% CI 9.6- 12.87; *P*=.007). The intervention district showed an 11.6 percentage-point increase in the average consistency of data among the delivery register and the integrated card compared with the comparator district (ATET 11.6; 95% CI 3.1-20.1; *P*=.04). Overall, the average consistency of data among IMRs in the intervention district was 11.6 times higher than that of the comparator district (ATET 11.6; 95% CI 4.2- 19; *P*=.03; Table 5).

Table 5. Effect of performance-based nonfinancial incentives on data consistency in individual medical records of institutional births, northwest Ethiopia, 2021.

Consistency	Intervention, mean (SD)	Comparator, mean (SD)	Intervention effect, ATET[a] (95% CI)[b]	*P* value
Delivery register vs integrated individual folder	62.2 (36.4)	51.1 (45.5)	11.2 (9.6-12.8)	.007
Delivery register vs integrated card	83.5 (19.9)	72.6 (26.0)	11.6 (3.1-20.1)	.04
Overall	80.2 (19.1)	69.4 (26.1)	11.6 (4.2-19)	.03

[a]ATET: average treatment effect on the treated.

[b]ATET estimates adjusted for covariates (presence of health information technology personnel, availability of data recording tools, availability of trained providers, supportive supervision from higher officials, existence of the Performance Monitoring Team [PMT], PMT per membership standard, monthly PMT meeting, monthly conducted lot quality assurance sampling, conducted root cause analysis, internal supervision, and availability of Health Management Information System guidelines).

Discussion

Principal Findings

This study evaluated the effect of PBNI on the quality of institutional birth data in northwest Ethiopia. PBNI improved both data completeness and consistency. The intervention district showed a 12% increase in data completeness compared with the comparator district, which showed a 1% decrease. Regarding data consistency, the intervention district improved by 18%, whereas the comparator district saw a 3% improvement. Controlling for other variables in the DID analysis, women’s

IMRs from the intervention district exhibited nearly 4 times higher data completeness and approximately 12 times greater data consistency than the comparator district.

Comparison With Prior Work

This study revealed a positive effect of PBNI on the data completeness and consistency of women's IMRs for institutional births. This finding aligns with that of previous studies that demonstrate the effectiveness of nonfinancial incentives in improving different aspects of health care quality. For instance, nonfinancial incentives have been proven to enhance the quality of medical error RCA in a US study [50]. Furthermore, studies from India and El Salvador [51,54,76] have shown an increase in the equitable and quality provision of maternal and child services. Nonfinancial incentives have also been reported to enhance quality consultations, according to studies from Tanzania [55,77]. The demonstrated effectiveness across contexts suggests the adaptability of nonfinancial incentives to improve the data quality and the quality of pregnancy-related services at health care facilities. These findings are particularly relevant for resource-limited settings where poor health care quality is associated with persistent mortality rates among mothers and children [3,5,78].

In this study, PBNI induced a greater extent of change compared with that in previous studies [51,54]. This difference may be due to differences in the incentive structures among the studies. In contrast to prior research that used team-based incentives, this study provided incentives at 3 levels: health facilities, departments, and individual health workers. Notably, the similarity between previous and current studies is apparent in the use of team-based incentives, reflected in this study's provision of incentives at the departmental level. Some earlier studies support the efficacy of team-based incentives in health care, emphasizing their role in fostering collective engagement [79-84]. Despite variations in the magnitude of the effect, the findings of this study do not contradict earlier research on the effectiveness of team-based incentives. Instead, the findings assert the potential for increased effectiveness by combining team-based, individual, and facility-level incentives.

According to this study, the effect of PBNI on data quality varies across the 3 women's records (integrated card, IIF, and delivery register). For example, although the integrated card saw a 3% increase in data completeness, women's folders increased by approximately 7%. These variations may suggest that the effectiveness of PBNI varies across health workers' professions. In Ethiopia, for instance, nonprofessional health workers are largely responsible for women's folder data, whereas midwives and other professionals are responsible for recording integrated card data [66]. Previous research in India has also demonstrated that the effectiveness of nonfinancial incentives varies depending on the health workers' professions, with frontline health workers experiencing a greater performance than supervisors [54]. Another study in northwest Ethiopia found a strong correlation between health worker motivation and their professional category [85]. These findings indicate the importance of recognizing the differences in the effectiveness of incentives and tailoring interventions to specific groups of health workers. Hence, policy makers and health care managers need to consider these variations when designing incentive programs, adopting a flexible approach that accounts for diverse roles and responsibilities.

This study reinforces the existing evidence that favors nonfinancial incentives in health care over financial incentives [86-90]. Concerns about financial incentives contradicting health care providers' intrinsic motivation to deliver quality care are widespread [86,90-100]. Therefore, this study suggests the practical use of nonfinancial incentives to enhance health care quality, especially in countries such as Ethiopia with limited financial capability. Prior studies in African countries have also indicated the importance of nonfinancial incentives in improving health care quality [44,101,102].

Policy and Research Implications

This study introduces PBNI as an effective intervention to improve the quality of institutional birth data. These findings underscore the potential of PBNI to complement established interventions to enhance RHIS performance, such as supportive supervision, mentorship, training, and feedback.

As this study evaluated the effect of PBNI on institutional birth data—a core indicator of maternal health care quality—the implications extend to broader RHIS data related to maternal and child services. These findings indicate the potential of PBNI to improve the quality of health services, which can contribute to maternal and child morbidity and mortality reduction. Hence, health care planners can consider adapting PBNI to improve the quality of maternal and child health services.

This study examined the effectiveness of PBNI in the context of health workers in health centers. Future studies are essential to understand the impact of PBNI on health staff across diverse settings, including health posts and hospitals.

Strengths and Limitations of the Study

One of the strengths of this study lies in its evaluation of the effect of PBNI on data quality within women's IMR in institutional births. Unlike most previous studies on RHIS data quality, which studied health facilities as the unit of analysis, this study delved into the individual level of data quality, which is essential to understanding how PBNI influences client-level service quality. In addition, we attempted to detect the minimum effect of the PBNI, using a sufficient study sample, and compared the intervention with comparator sites, increasing the robustness of the findings. Furthermore, to establish a causal effect of PBNI, the study used DID analysis, a recognized causal analysis technique in nonrandomized studies. Nevertheless, it is essential to recognize the limitations of this study. First, the retrospective design prevented randomization in the selection of the study participants. Second, interviewer bias is possible during the completeness and consistency assessments, as data collectors judged these aspects despite training to reduce bias. Although we attempted to disentangle the effect of PBNI from other potential factors, unmeasured confounders may still exist. Moreover, the security issues in Northern Ethiopia might have disrupted the effectiveness of the PBNI, as the intervention coincided with security problems in the adjacent regions.

Conclusions

This study shows that PBNI improves institutional birth data quality, as demonstrated by enhanced completeness and consistency. The effectiveness of PBNI can be extended to enhancing comparable RHIS data in maternal and child care and improving service quality at health care facilities. Health care planners can consider PBNI to enhance the quality of maternal and child health services in health care facilities. Future studies are essential to understand the impact of PBNI in diverse health care settings.

Acknowledgments

This work was financially supported by the Doris Duke Charitable Foundation (grant 2017187). The mission of the Doris Duke Charitable Foundation is to improve the quality of people's lives through grants supporting the performing arts, environmental conservation, medical research, and child well-being and through the preservation of the cultural and environmental legacy of Doris Duke's properties. The authors would like to express their gratitude to the supervisors, data collectors, and health office departments of the Central Gondar Zone, Wegera district, and Tach-armachiho district.

Authors' Contributions

BKT conceived and designed the study, performed the data collection, analyzed and interpreted the data, and drafted the manuscript. LDG, AA, SAM, JK, MKG, and BT analyzed and interpreted the data, and contributed to writing the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Average data completeness across individual medical records of institutional births during pre- and postintervention periods, northwest Ethiopia, 2021.
[PNG File , 35 KB - medinform_v12i1e54278_app1.png]

References

1. New global targets to prevent maternal deaths. World Health Organization. 2021 Oct 05. URL: https://www.who.int/news/item/05-10-2021-new-global-targets-to-prevent-maternal-deaths [accessed 2023-05-05]
2. Say L, Chou D, Gemmill A, Tunçalp Ö, Moller AB, Daniels J, et al. Global causes of maternal death: a WHO systematic analysis. Lancet Glob Health 2014 Jun;2(6):e323-e333 [FREE Full text] [doi: 10.1016/S2214-109X(14)70227-X] [Medline: 25103301]
3. Maternal mortality. World Health Organization. 2023 Feb 22. URL: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality [accessed 2022-11-17]
4. Sk MI, Paswan B, Anand A, Mondal NA. Praying until death: revisiting three delays model to contextualize the socio-cultural factors associated with maternal deaths in a region with high prevalence of eclampsia in India. BMC Pregnancy Childbirth 2019 Aug 28;19(1):314 [FREE Full text] [doi: 10.1186/s12884-019-2458-5] [Medline: 31455258]
5. Trends in maternal mortality 2000 to 2020: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/population division. World Health Organization. 2023 Feb 23. URL: https://www.who.int/publications/i/item/9789240068759 [accessed 2024-03-15]
6. Tunçalp , Were WM, MacLennan C, Oladapo OT, Gülmezoglu AM, Bahl R, et al. Quality of care for pregnant women and newborns-the WHO vision. BJOG 2015 Jul;122(8):1045-1049 [FREE Full text] [doi: 10.1111/1471-0528.13451] [Medline: 25929823]
7. van den Broek NR, Graham WJ. Quality of care for maternal and newborn health: the neglected agenda. BJOG 2009 Oct;116 Suppl 1:18-21. [doi: 10.1111/j.1471-0528.2009.02333.x] [Medline: 19740165]
8. SDG Target 3.1 Reduce the global maternal mortality ratio to less than 70 per 100 000 live births. World Health Organization. URL: https://www.who.int/data/gho/data/themes/topics/sdg-target-3-1-maternal-mortality [accessed 2023-02-09]
9. Wagenaar BH, Sherr K, Fernandes Q, Wagenaar AC. Using routine health information systems for well-designed health evaluations in low- and middle-income countries. Health Policy Plan 2016 Feb;31(1):129-135 [FREE Full text] [doi: 10.1093/heapol/czv029] [Medline: 25887561]
10. Shamba D, Day LT, Zaman SB, Sunny AK, Tarimo MN, Peven K, et al. Barriers and enablers to routine register data collection for newborns and mothers: EN-BIRTH multi-country validation study. BMC Pregnancy Childbirth 2021 Mar 26;21(Suppl 1):233 [FREE Full text] [doi: 10.1186/s12884-020-03517-3] [Medline: 33765963]
11. Framework and standards for country health information systems, 2nd edition. World Health Organization. 2023 Apr 24. URL: https://www.who.int/publications/i/item/9789241595940 [accessed 2024-03-15]

12. Standards for improving quality of maternal and newborn care in health facilities. World Health Organization. 2016. URL: https://pesquisa.bvsalud.org/portal/resource/pt/per-3087?lang=en [accessed 2022-11-08]
13. Dehnavieh R, Haghdoost A, Khosravi A, Hoseinabadi F, Rahimi H, Poursheikhali A, et al. The District Health Information System (DHIS2): a literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries. Health Inf Manag 2019 May;48(2):62-75. [doi: 10.1177/1833358318777713] [Medline: 29898604]
14. Hung YW, Hoxha K, Irwin BR, Law MR, Grépin KA. Using routine health information data for research in low- and middle-income countries: a systematic review. BMC Health Serv Res 2020 Aug 25;20(1):790 [FREE Full text] [doi: 10.1186/s12913-020-05660-1] [Medline: 32843033]
15. Shama AT, Roba HS, Abaerei AA, Gebremeskel TG, Baraki N. Assessment of quality of routine health information system data and associated factors among departments in public health facilities of Harari region, Ethiopia. BMC Med Inform Decis Mak 2021 Oct 19;21(1):287 [FREE Full text] [doi: 10.1186/s12911-021-01651-2] [Medline: 34666753]
16. Lippeveld T. Routine health facility and community information systems: creating an information use culture. Glob Health Sci Pract 2017 Sep 28;5(3):338-340 [FREE Full text] [doi: 10.9745/GHSP-D-17-00319] [Medline: 28963169]
17. O'Hagan R, Marx MA, Finnegan KE, Naphini P, Ng'ambi K, Laija K, et al. National assessment of data quality and associated systems-level factors in Malawi. Glob Health Sci Pract 2017 Sep 28;5(3):367-381 [FREE Full text] [doi: 10.9745/GHSP-D-17-00177] [Medline: 28963173]
18. Mwinnyaa G, Hazel E, Maïga A, Amouzou A. Estimating population-based coverage of reproductive, maternal, newborn, and child health (RMNCH) interventions from health management information systems: a comprehensive review. BMC Health Serv Res 2021 Oct 25;21(Suppl 2):1083 [FREE Full text] [doi: 10.1186/s12913-021-06995-z] [Medline: 34689787]
19. Begum T, Khan SM, Adamou B, Ferdous J, Parvez MM, Islam MS, et al. Perceptions and experiences with district health information system software to collect and utilize health data in Bangladesh: a qualitative exploratory study. BMC Health Serv Res 2020 May 26;20(1):465 [FREE Full text] [doi: 10.1186/s12913-020-05322-2] [Medline: 32456706]
20. Day LT, Ruysen H, Gordeev VS, Gore-Langton GR, Boggs D, Cousens S, et al. "Every Newborn-BIRTH" protocol: observational study validating indicators for coverage and quality of maternal and newborn health care in Bangladesh, Nepal and Tanzania. J Glob Health 2019 Jun;9(1):010902 [FREE Full text] [doi: 10.7189/jogh.09.010902] [Medline: 30863542]
21. Alaro T, Sisay S, Samuel S. Implementation level of health management information system program in governmental hospitals of Ethiopia. Int J Intell Inf Syst 2019 Apr;8(2):52-57. [doi: 10.11648/j.ijiis.20190802.13]
22. HMIS indicator reference guide. International Institute for Primary Health Care Ethiopia. 2017. URL: http://repository.iifphc.org/handle/123456789/392 [accessed 2022-11-17]
23. Lemma S, Janson A, Persson L, Wickremasinghe D, Källestål C. Improving quality and use of routine health information system data in low- and middle-income countries: a scoping review. PLoS One 2020 Oct 8;15(10):e0239683 [FREE Full text] [doi: 10.1371/journal.pone.0239683] [Medline: 33031406]
24. Alemu MB, Atnafu A, Gebremedhin T, Endehabtu BF, Asressie M, Tilahun B. Outcome evaluation of capacity building and mentorship partnership (CBMP) program on data quality in the public health facilities of Amhara National Regional State, Ethiopia: a quasi-experimental evaluation. BMC Health Serv Res 2021 Oct 05;21(1):1054 [FREE Full text] [doi: 10.1186/s12913-021-07063-2] [Medline: 34610844]
25. Kanfe SG, Endehabtu BF, Ahmed MH, Mengestie ND, Tilahun B. Commitment levels of health care providers in using the district health information system and the associated factors for decision making in resource-limited settings: cross-sectional survey study. JMIR Med Inform 2021 Mar 04;9(3):e23951 [FREE Full text] [doi: 10.2196/23951] [Medline: 33661133]
26. Adane A, Adege TM, Ahmed MM, Anteneh HA, Ayalew ES, Berhanu D, et al. Routine health management information system data in Ethiopia: consistency, trends, and challenges. Glob Health Action 2021 Jan 01;14(1):1868961 [FREE Full text] [doi: 10.1080/16549716.2020.1868961] [Medline: 33446081]
27. Endriyas M, Kawza A, Alano A, Lemango F. Quality of medical records in public health facilities: a case of Southern Ethiopia, resource limited setting. Health Informatics J 2022;28(3):14604582221112853 [FREE Full text] [doi: 10.1177/14604582221112853] [Medline: 35793497]
28. Arsenault C, Yakob B, Kassa M, Dinsa G, Verguet S. Using health management information system data: case study and verification of institutional deliveries in Ethiopia. BMJ Glob Health 2021 Aug;6(8):e006216 [FREE Full text] [doi: 10.1136/bmjgh-2021-006216] [Medline: 34426404]
29. Hoxha K, Hung YW, Irwin BR, Grépin KA. Understanding the challenges associated with the use of data from routine health information systems in low- and middle-income countries: a systematic review. Health Inf Manag 2022 Sep;51(3):135-148. [doi: 10.1177/1833358320928729] [Medline: 32602368]
30. Straus S, Tetroe J, Graham ID. Knowledge Translation in Health Care: Moving from Evidence to Practice. Hoboken, NJ: Wiley; 2009.
31. Godin G, Bélanger-Gravel A, Eccles M, Grimshaw J. Healthcare professionals' intentions and behaviours: a systematic review of studies based on social cognitive theories. Implement Sci 2008 Jul 16;3:36 [FREE Full text] [doi: 10.1186/1748-5908-3-36] [Medline: 18631386]

32. Eccles MP, Johnston M, Hrisos S, Francis J, Grimshaw J, Steen N, et al. Translating clinicians' beliefs into implementation interventions (TRACII): a protocol for an intervention modeling experiment to change clinicians' intentions to implement evidence-based practice. Implement Sci 2007 Aug 16;2:27 [FREE Full text] [doi: 10.1186/1748-5908-2-27] [Medline: 17705824]
33. Flodgren G, Eccles MP, Shepperd S, Scott A, Parmelli E, Beyer FR. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. Cochrane Database Syst Rev 2011 Jul 06;2011(7):CD009255 [FREE Full text] [doi: 10.1002/14651858.CD009255] [Medline: 21735443]
34. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev 2015;9(3):323-344 [FREE Full text] [doi: 10.1080/17437199.2014.941722] [Medline: 25104107]
35. Eccles M, Grimshaw J, Walker A, Johnston M, Pitts N. Changing the behavior of healthcare professionals: the use of theory in promoting the uptake of research findings. J Clin Epidemiol 2005 Feb;58(2):107-112. [doi: 10.1016/j.jclinepi.2004.09.002] [Medline: 15680740]
36. Saint-Lary O, Plu I, Naiditch M. Ethical issues raised by the introduction of payment for performance in France. J Med Ethics 2012 Aug;38(8):485-491 [FREE Full text] [doi: 10.1136/medethics-2011-100159] [Medline: 22493186]
37. Basinga P, Gertler PJ, Binagwaho A, Soucat AL, Sturdy J, Vermeersch CM. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. Lancet 2011 Apr 23;377(9775):1421-1428 [FREE Full text] [doi: 10.1016/S0140-6736(11)60177-3] [Medline: 21515164]
38. Eichler R, Agarwal K, Askew I, Iriarte E, Morgan L, Watson J. Performance-based incentives to improve health status of mothers and newborns: what does the evidence show? J Health Popul Nutr 2013 Dec;31(4 Suppl 2):36-47. [Medline: 24992802]
39. Mendelson A, Kondo K, Damberg C, Low A, Motúapuaka M, Freeman M, et al. The effects of pay-for-performance programs on health, health care use, and processes of care: a systematic review. Ann Intern Med 2017 Mar 07;166(5):341-353 [FREE Full text] [doi: 10.7326/M16-1881] [Medline: 28114600]
40. Diaconu K, Falconer J, Verbel A, Fretheim A, Witter S. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. Cochrane Database Syst Rev 2021 May 05;5(5):CD007899 [FREE Full text] [doi: 10.1002/14651858.CD007899.pub3] [Medline: 33951190]
41. Bufalino V, Peterson ED, Burke GL, LaBresh KA, Jones DW, Faxon DP, et al. Payment for quality: guiding principles and recommendations: principles and recommendations from the American Heart Association's Reimbursement, Coverage, and Access Policy Development Workgroup. Circulation 2006 Feb 28;113(8):1151-1154. [doi: 10.1161/CIRCULATIONAHA.105.171760] [Medline: 16401766]
42. Soeters R, Peerenboom PB, Mushagalusa P, Kimanuka C. Performance-based financing experiment improved health care in the Democratic Republic of Congo. Health Aff (Millwood) 2011 Aug;30(8):1518-1527. [doi: 10.1377/hlthaff.2009.0019] [Medline: 21821568]
43. Huntington D, Zaky HH, Shawky S, Fattah FA, El-Hadary E. Impact of a service provider incentive payment scheme on quality of reproductive and child-health services in Egypt. J Health Popul Nutr 2010 Jun;28(3):273-280 [FREE Full text] [doi: 10.3329/jhpn.v28i3.5556] [Medline: 20635638]
44. Lagarde M, Huicho L, Papanicolas I. Motivating provision of high quality care: it is not all about the money. BMJ 2019 Sep 23;366:l5210 [FREE Full text] [doi: 10.1136/bmj.l5210] [Medline: 31548200]
45. Leonard KL, Masatu MC. Professionalism and the know-do gap: exploring intrinsic motivation among health workers in Tanzania. Health Econ 2010 Dec;19(12):1461-1477. [doi: 10.1002/hec.1564] [Medline: 19960481]
46. Ochenge NC, Susan W. Role of reward systems in employee motivation in Kenyan deposit taking micro finance institutions: a case study of Faulu Kenya. Int J Soc Sci Manag Entrep 2014;1(2):203-220.
47. Burgess S, Metcalfe R, Sadoff S. Understanding the response to financial and non-financial incentives in education: field experimental evidence using high-stakes assessments. Econ Educ Rev 2021 Dec;85:102195 [FREE Full text] [doi: 10.1016/j.econedurev.2021.102195]
48. Ashraf N, Bandiera O, Jack BK. No margin, no mission? A field experiment on incentives for public service delivery. J Public Econ 2014 Dec;120:1-17 [FREE Full text] [doi: 10.1016/j.jpubeco.2014.06.014]
49. Ashraf N, Bandiera O, Lee SS. Do-gooders and go-getters: career incentives, selection, and performance in public service delivery. Harvard Business School. 2015 Mar. URL: https://www.hbs.edu/faculty/Pages/item.aspx?num=46043 [accessed 2024-03-25]
50. Bagian JP, King BJ, Mills PD, McKnight SD. Improving RCA performance: the Cornerstone Award and the power of positive reinforcement. BMJ Qual Saf 2011 Nov;20(11):974-982. [doi: 10.1136/bmjqs.2010.049585] [Medline: 21775506]
51. Bernal P, Martinez S. In-kind incentives and health worker performance: experimental evidence from El Salvador. J Health Econ 2020 Mar;70:102267 [FREE Full text] [doi: 10.1016/j.jhealeco.2019.102267] [Medline: 32028090]
52. Bufalino V, Peterson ED, Krumholz HM, Burke GL, LaBresh KA, Jones DW, et al. Nonfinancial incentives for quality: a policy statement from the American Heart Association. Circulation 2007 Jan 23;115(3):398-401. [doi: 10.1161/CIRCULATIONAHA.106.180202] [Medline: 17179024]

53. Cacace M, Geraedts M, Berger E. Public reporting as a quality strategy. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies. Copenhagen, Denmark: European Observatory on Health Systems and Policies; 2019.
54. Grant C, Nawal D, Guntur SM, Kumar M, Chaudhuri I, Galavotti C, et al. 'We pledge to improve the health of our entire community': improving health worker motivation and performance in Bihar, India through teamwork, recognition, and non-financial incentives. PLoS One 2018 Aug 30;13(8):e0203265 [FREE Full text] [doi: 10.1371/journal.pone.0203265] [Medline: 30161213]
55. Brock JM, Lange A, Leonard KL. Giving and promising gifts: experimental evidence on reciprocity from the field. J Health Econ 2018 Mar;58:188-201 [FREE Full text] [doi: 10.1016/j.jhealeco.2018.02.007] [Medline: 29524793]
56. Population size of towns by sex, region, zone and Weredas as of July 2021. Ethiopian Statistical Service. URL: https://www.statsethiopia.gov.et/wp-content/uploads/2020/08/Population-of-Towns-as-of-July-2021.pdf [accessed 2024-03-25]
57. National guideline for family planning services in Ethiopia. Ministry of Health, Federal Democratic Republic of Ethiopia. 2011 Feb. URL: https://pdf4pro.com/view/national-family-planning-guideline-phe-ethiopia-1876ed.html [accessed 2023-10-20]
58. Asmamaw A, Tesfahun H, Berhanu FE, Lemma DG, Adane M, Teklehaymanot G, et al. Implementation outcomes of performance based non- financial incentive: using RE-AIM framework. Ethiop J Health Dev 2023;37(1):1-10.
59. Capacity building and mentorship program (CBMP). eHealthlab Ethiopia. URL: https://ehealthlab.org/cbmp/ [accessed 2024-03-25]
60. Chanyalew MA, Yitayal M, Atnafu A, Mengiste SA, Tilahun B. The effectiveness of the capacity building and mentorship program in improving evidence-based decision-making in the Amhara Region, Northwest Ethiopia: difference-in-differences study. JMIR Med Inform 2022 Apr 22;10(4):e30518 [FREE Full text] [doi: 10.2196/30518] [Medline: 35451990]
61. Tilahun B, Endehabtu BF, Hailemariam T, Derseh Gezie L, Mamuye A, Gebrehiwot T, et al. Effectiveness of performance-based non-financial incentive for improved health data quality and information use at primary health care units, northwest Ethiopia. Ethiop J Health Dev 2023 Nov 16;37(1):1-9. [doi: 10.20372/ejhd.v37i1.5839]
62. Witter S, Bertone MP, Diaconu K, Bornemisza O. Performance-based financing versus "unconditional" direct facility financing - false dichotomy? Health Syst Reform 2021 Jan 01;7(1):e2006121. [doi: 10.1080/23288604.2021.2006121] [Medline: 34874806]
63. Quentin W, Partanen VM, Brownwood I, Klazinga N. Measuring healthcare quality. In: Busse R, Klazinga N, Panteli D, Quentin W, editors. Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies. Copenhagen, Denmark: European Observatory on Health Systems and Policies; 2019.
64. Gezie LD, Endehabtu BF, Hailemariam T, Atnafu A, Mamuye A, Mohamed M, et al. Barriers and facilitators of implementing performance- based non- financial incentives to improve data quality and use: using a consolidated framework for implementation research. Ethiop J Health Dev 2023;37(1):1-10. [doi: 10.20372/ejhd.v37i1.5838]
65. Amare G, Minyihun A, Atnafu A, Endehabtu BF, Derseh L, Hailemariam T, et al. Cost-effectiveness of performance-based non-financial incentive (PBNI) intervention to improve health information system performance at Wogera district in northwest Ethiopia. Ethiop J Health Dev 2023 Nov 16;37(1):1-11. [doi: 10.20372/ejhd.v37i1.5837]
66. Ethiopian health management information system: data recording and reporting procedures manual. Federal Ministry of Health Ethiopia. 2017 Jul. URL: http://dataverse.nipn.ephi.gov.et/bitstream/handle/123456789/293/HMIS%20Recording%20and%20Reporting%20Procedures.pdf?sequence=1 [accessed 2022-11-17]
67. Zozus MN, Hammond WE, Green BB, Kahn MG, Richesson RL, Rusincovitch SA, et al. Assessing data quality for healthcare systems data used in clinical research. NIH Pragmatic Trials Collaboratory. 2014. URL: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Assessing-data-quality_V1%200.pdf [accessed 2023-10-20]
68. Data quality assurance: module 3: site assessment of data quality: data verification and system assessment. World Health Organization. 2023 Jan 30. URL: https://www.who.int/publications/i/item/9789240049123 [accessed 2023-10-20]
69. Maïga A, Jiwani SS, Mutua MK, Porth TA, Taylor CM, Asiki G, et al. Generating statistics from health facility data: the state of routine health information systems in Eastern and Southern Africa. BMJ Glob Health 2019 Sep 29;4(5):e001849 [FREE Full text] [doi: 10.1136/bmjgh-2019-001849] [Medline: 31637032]
70. PRISM: performance of routine information system management series. Measure Evaluation. URL: https://tinyurl.com/bdhap69j [accessed 2024-03-17]
71. Aqil A, Lippeveld T, Hozumi D. PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. Health Policy Plan 2009 May;24(3):217-228 [FREE Full text] [doi: 10.1093/heapol/czp010] [Medline: 19304786]
72. Health data quality training module participant manual. Federal Democratic Republic Of Ethiopia, Ministry of Health. 2018. URL: https://tinyurl.com/3rc5uts3 [accessed 2024-02-11]
73. Taye BK, Gezie LD, Atnafu A, Mengiste SA, Tilahun B. Data completeness and consistency in individual medical records of institutional births: retrospective crossectional study from Northwest Ethiopia, 2022. BMC Health Serv Res 2023 Oct 31;23(1):1189 [FREE Full text] [doi: 10.1186/s12913-023-10127-0] [Medline: 37907881]
74. Luedicke J. Difference-in-differences estimation using Stata. Stata Users Group. 2022. URL: https://ideas.repec.org/p/boc/dsug22/06.html [accessed 2024-02-11]

75. Introduction to difference-in-differences estimation. In: Causal Inference and Treatment-Effects Estimation Reference Manual. College Station, TX: Stata Press; 2023.
76. Carmichael SL, Mehta K, Raheel H, Srikantiah S, Chaudhuri I, Trehan S, et al. Effects of team-based goals and non-monetary incentives on front-line health worker performance and maternal health behaviours: a cluster randomised controlled trial in Bihar, India. BMJ Glob Health 2019 Aug 26;4(4):e001146 [FREE Full text] [doi: 10.1136/bmjgh-2018-001146] [Medline: 31543982]
77. Brock MJ, Lange A, Leonard KL. Generosity norms and intrinsic motivation in health care provision: evidence from the laboratory and the field. European Bank for Reconstruction and Development, Office of the Chief Economist. 2012. URL: https://ideas.repec.org/p/ebd/wpaper/147.html [accessed 2024-02-11]
78. Newborn mortality. World Health Organization. 2024 Mar 14. URL: https://www.who.int/news-room/fact-sheets/detail/levels-and-trends-in-child-mortality-report-2021 [accessed 2024-03-25]
79. Mbindyo P, Gilson L, Blaauw D, English M. Contextual influences on health worker motivation in district hospitals in Kenya. Implement Sci 2009 Jul 23;4:43 [FREE Full text] [doi: 10.1186/1748-5908-4-43] [Medline: 19627590]
80. Benjamin L, Cotté FE, Philippe C, Mercier F, Bachelot T, Vidal-Trécan G. Physicians' preferences for prescribing oral and intravenous anticancer drugs: a discrete choice experiment. Eur J Cancer 2012 Apr;48(6):912-920. [doi: 10.1016/j.ejca.2011.09.019] [Medline: 22033327]
81. Sharma R, Webster P, Bhattacharyya S. Factors affecting the performance of community health workers in India: a multi-stakeholder perspective. Glob Health Action 2014 Oct 13;7:25352 [FREE Full text] [doi: 10.3402/gha.v7.25352] [Medline: 25319596]
82. Kivimäki M, Vanhala A, Pentti J, Länsisalmi H, Virtanen M, Elovainio M, et al. Team climate, intention to leave and turnover among hospital employees: prospective cohort study. BMC Health Serv Res 2007 Oct 23;7:170 [FREE Full text] [doi: 10.1186/1472-6963-7-170] [Medline: 17956609]
83. Borkum E, Rangarajan A, Rotz D, Sridharan S, Sethi S, Manorajini M. Evaluation of the team-based goals and performance-based incentives (TBGI) innovation in Bihar. Mathematica Policy Research Reports. URL: https://ideas.repec.org/p/mpr/mprres/d8e1097122ff47a6bf42580c82677834.html [accessed 2024-02-11]
84. Lee TH, Bothe A, Steele GD. How Geisinger structures its physicians' compensation to support improvements in quality, efficiency, and volume. Health Aff (Millwood) 2012 Sep;31(9):2068-2073. [doi: 10.1377/hlthaff.2011.0940] [Medline: 22949457]
85. Weldegebriel Z, Ejigu Y, Weldegebreal F, Woldie M. Motivation of health workers and associated factors in public hospitals of West Amhara, Northwest Ethiopia. Patient Prefer Adherence 2016 Feb 15;10:159-169 [FREE Full text] [doi: 10.2147/PPA.S90323] [Medline: 26929608]
86. Lee TH. Financial versus non-financial incentives for improving patient experience. J Patient Exp 2015 May;2(1):4-6 [FREE Full text] [doi: 10.1177/237437431500200102] [Medline: 28725809]
87. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ 2016 Jan;50(1):36-60. [doi: 10.1111/medu.12817] [Medline: 26695465]
88. Lagarde M, Blaauw D. Pro-social preferences and self-selection into jobs: evidence from South African nurses. J Econ Behav Organ 2014 Nov;107(A):136-152 [FREE Full text] [doi: 10.1016/j.jebo.2014.09.004]
89. Ashraf N, Bandiera O. Altruistic capital. Am Econ Rev 2017 May;107(5):70-75. [doi: 10.1257/aer.p20171097]
90. Attema AE, Galizzi MM, Groß M, Hennig-Schmidt H, Karay Y, L'Haridon O, et al. The formation of physician altruism. J Health Econ 2023 Jan;87:102716 [FREE Full text] [doi: 10.1016/j.jhealeco.2022.102716] [Medline: 36603361]
91. Khullar D, Wolfson D, Casalino LP. Professionalism, performance, and the future of physician incentives. JAMA 2018 Dec 18;320(23):2419-2420. [doi: 10.1001/jama.2018.17719] [Medline: 30476944]
92. Scott A, Sivey P, Ait Ouakrim D, Willenberg L, Naccarella L, Furler J, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database Syst Rev 2011 Sep 07(9):CD008451. [doi: 10.1002/14651858.CD008451.pub2] [Medline: 21901722]
93. Witter S, Fretheim A, Kessy FL, Lindahl AK. Paying for performance to improve the delivery of health interventions in low- and middle-income countries . Cochrane Database Syst Rev 2012 Feb 15(2):CD007899. [doi: 10.1002/14651858.CD007899.pub2] [Medline: 22336833]
94. Berwick DM. Era 3 for medicine and health care. JAMA 2016 Apr 05;315(13):1329-1330. [doi: 10.1001/jama.2016.1509] [Medline: 26940610]
95. Delfgaauw J. Dedicated doctors: public and private provision of health care with altruistic physicians. SSRN Preprint posted online February 5, 2007. [doi: 10.2139/ssrn.958693]
96. Liu T, Ma CT. Health insurance, treatment plan, and delegation to altruistic physician. J Econ Behav Org 2013 Jan;85:79-96 [FREE Full text] [doi: 10.1016/j.jebo.2012.11.002]
97. Mannion R, Davies HT. Payment for performance in health care. BMJ 2008 Feb 09;336(7639):306-308 [FREE Full text] [doi: 10.1136/bmj.39463.454815.94] [Medline: 18258966]
98. Chen LM, Epstein AM, Orav EJ, Filice CE, Samson LW, Joynt Maddox KE. Association of practice-level social and medical risk with performance in the medicare physician value-based payment modifier program. JAMA 2017 Aug 01;318(5):453-461 [FREE Full text] [doi: 10.1001/jama.2017.9643] [Medline: 28763549]

99. Heath I, Hippisley-Cox J, Smeeth L. Measuring performance and missing the point? BMJ 2007 Nov 24;335(7629):1075-1076 [FREE Full text] [doi: 10.1136/bmj.39377.387373.AD] [Medline: 18033930]
100. McDonald R, Harrison S, Checkland K, Campbell SM, Roland M. Impact of financial incentives on clinical autonomy and internal motivation in primary care: ethnographic study. BMJ 2007 Jun 30;334(7608):1357 [FREE Full text] [doi: 10.1136/bmj.39238.890810.BE] [Medline: 17580318]
101. Borghi J, Little R, Binyaruka P, Patouillard E, Kuwawenaruwa A. In Tanzania, the many costs of pay-for-performance leave open to debate whether the strategy is cost-effective. Health Aff (Millwood) 2015 Mar;34(3):406-414. [doi: 10.1377/hlthaff.2014.0608] [Medline: 25732490]
102. Mathauer I, Imhoff I. Health worker motivation in Africa: the role of non-financial incentives and human resource management tools. Hum Resour Health 2006 Aug 29;4(1):24 [FREE Full text] [doi: 10.1186/1478-4491-4-24] [Medline: 16939644]

Abbreviations

ATET: average treatment effect on the treated
CBMP: Capacity Building and Mentorship Program
DID: difference-in-differences
HIT: health IT
HMIS: Health Management Information System
IIF: integrated individual folder
IMR: individual medical record
LQAS: lot quality assurance sampling
MRN: medical record number
MRU: Medical Record Unit
PBNI: performance-based nonfinancial incentives
PMT: Performance Monitoring Team
RCA: root cause analysis
RHIS: Routine Health Information System

Edited by C Perrin; submitted 03.11.23; peer-reviewed by T Wonde, T Tefera; comments to author 11.01.24; revised version received 20.01.24; accepted 05.02.24; published 05.04.24.

Please cite as:
Taye BK, Gezie LD, Atnafu A, Mengiste SA, Kaasbøll J, Gullslett MK, Tilahun B
Effect of Performance-Based Nonfinancial Incentives on Data Quality in Individual Medical Records of Institutional Births: Quasi-Experimental Study
JMIR Med Inform 2024;12:e54278
URL: https://medinform.jmir.org/2024/1/e54278
doi:10.2196/54278
PMID:38578684

Data Flow Construction and Quality Evaluation of Electronic Source Data in Clinical Trials: Pilot Study Based on Hospital Electronic Medical Records in China

Yannan Yuan[1], MS; Yun Mei[2], MS; Shuhua Zhao[1], MS; Shenglong Dai[3], MS; Xiaohong Liu[1], MS; Xiaojing Sun[3], MA; Zhiying Fu[1], MS; Liheng Zhou[3], MS; Jie Ai[2], MS; Liheng Ma[3], MD; Min Jiang[4], MS

1

2

3

4

Corresponding Author:
Min Jiang, MS

Abstract

Background: The traditional clinical trial data collection process requires a clinical research coordinator who is authorized by the investigators to read from the hospital's electronic medical record. Using electronic source data opens a new path to extract patients' data from electronic health records (EHRs) and transfer them directly to an electronic data capture (EDC) system; this method is often referred to as eSource. eSource technology in a clinical trial data flow can improve data quality without compromising timeliness. At the same time, improved data collection efficiency reduces clinical trial costs.

Objective: This study aims to explore how to extract clinical trial–related data from hospital EHR systems, transform the data into a format required by the EDC system, and transfer it into sponsors' environments, and to evaluate the transferred data sets to validate the availability, completeness, and accuracy of building an eSource dataflow.

Methods: A prospective clinical trial study registered on the Drug Clinical Trial Registration and Information Disclosure Platform was selected, and the following data modules were extracted from the structured data of 4 case report forms: demographics, vital signs, local laboratory data, and concomitant medications. The extracted data was mapped and transformed, deidentified, and transferred to the sponsor's environment. Data validation was performed based on availability, completeness, and accuracy.

Results: In a secure and controlled data environment, clinical trial data was successfully transferred from a hospital EHR to the sponsor's environment with 100% transcriptional accuracy, but the availability and completeness of the data could be improved.

Conclusions: Data availability was low due to some required fields in the EDC system not being available directly in the EHR. Some data is also still in an unstructured or paper-based format. The top-level design of the eSource technology and the construction of hospital electronic data standards should help lay a foundation for a full electronic data flow from EHRs to EDC systems in the future.

(JMIR Med Inform 2024;12:e52934) doi:10.2196/52934

KEYWORDS

clinical trials; electronic source data; EHRs; electronic data capture systems; data quality; electronic health records

Introduction

Source data are the original records from clinical trials or all information recorded on certified copies, including clinical findings, observations, and records of other relevant activities necessary for the reconstruction and evaluation of the trial [1]. Electronic source data are data initially recorded in an electronic format (electronic source data or eSource) [2,3].

The traditional clinical trial data collection process requires a clinical research coordinator (CRC) who is authorized by the investigators to read from the hospital's electronic medical record and other clinical trial–related data from the hospital information system and then manually enter the patient's data into the electronic data capture (EDC) system. After data entry, the clinical research associate visits the site to perform source data verification and source data review. The drawbacks of collecting data by manual transcription are that data quality and timeliness cannot be guaranteed and that it is a waste of human and material resources. Using electronic source data opens a new path to extract patients' data from electronic health records (EHRs) and transfer it directly to EDC systems (often the method is referred to as eSource) [4]. eSource technology in a clinical trial data flow can improve data quality without

compromising timeliness [5]. At the same time, improved data collection efficiency reduces clinical trial costs [6].

eSource can be divided into two levels. The first level is to enable the hospital information system to obtain complete data sets; the second level is to allow direct data transfer to EDC systems based on the clinical trial patients' electronic data in hospitals to avoid the electronic data being transcribed manually again, which is the core purpose of eSource [7]. This project will explore the use of eSource technology to extract clinical trial data from EHRs, send it to the sponsor data environment, and discuss the issues and challenges occurring in its application process.

Methods

Ethics Approval

This study was approved by the Ethics Committee and Human Genetic Resource Administration of China (2020YW135). During the ethical review process, the most significant challenges were patients' informed consent, privacy protection, and data security. The B7461024 Informed Consent Form (Version 4) states that "interested parties may use subjects' personal information to improve the quality, design, and safety of this and other studies," and "Is my personal information likely to be used in other studies? Your coded information may be used to advance scientific research and public health in other projects conducted in future." This project is an exploration of using electronic source data technology instead of traditional manual transcription in the process of transferring data from hospital EHRs to EDC systems, which will improve the data quality of clinical trials and will improve the data flow in the future. Therefore, this project is within the scope of the informed consent form for study B7461024, which was approved by the ethics committee after clarification.

Project Information

This project was conducted from December 15, 2020, to November 19, 2021, which was before China's personal information protection law and data security law were introduced. The data for this project were obtained from an ongoing phase 2, multicenter, open-label, dual-cohort study to evaluate the efficacy and safety of Lorlatinib (pf-06463922) monotherapy in anaplastic lymphoma kinase (ALK) inhibitor–treated locally advanced or metastatic ALK-positive non–small cell lung cancer patients in China (B7461024), registered by the sponsor on the Drug Clinical Trials Registration and Disclosure Platform (CTR20181867). The data extraction involved 4 case report form (CRF) data modules: demographics, concomitant medication, local lab, and vital signs, which were collected in the following ways:

- Demographics: Originally entered directly into the hospital EHR then manually transcribed by the CRC to the sponsor's EDC system
- Local lab: Laboratory data collected by the hospital laboratory information management system (LIMS) and then manually transcribed by the CRC into the EDC system
- Vital signs: Hospital uses paper-based tracking form provided by the sponsor to record patients' vital signs and investigators transcribe the vital signs data into the hospital medical record
- Concomitant medication: Similar to vital signs, hospital uses the paper tracking form provided by the sponsor to record the adverse reactions and concomitant medication; investigator might also transfer the concomitant medication data into the hospital EHR, but there was no mandatory requirement to transfer these data into patients' medical records

All information was collected from 6 patients in a total of 29 fields (Textbox 1).

Textbox 1. Data collection fields.

Demographics

- Subject ID
- Date of birth
- Sex
- Ethnicity
- Race
- Age

Concomitant medication

- Combined drug name
- Whether for the treatment of adverse reactions
- Adverse event number
- Combined drug start date
- Combined drug end date
- Currently still in use

Vital signs

- Date of vital signs collection
- Weight
- Weight unit
- Body temperature
- Height
- Height unit
- Location of temperature measurement
- Systolic blood pressure
- Diastolic blood pressure
- Pulse

Local lab

- Laboratory inspection name
- Laboratory name and address
- Sponsor number
- Laboratory number
- Incomplete laboratory inspection
- Sample collection data
- Inspection results

Data Process Workflow

Overview

The study chosen in our project used the traditional manual data entry method to transcribe patients' CRF data into the EDC system. This project proposes testing the acquisition of data directly from the hospital EHR, deidentification of the patients' electronic data on the hospital medical data intelligence platform, mapping and transforming the data based on the sponsor's EDC data standard, and transferring the data into the sponsor's environment. The data was transferred from the hospital to the sponsor's data environment and compared to data that was captured by traditional manual entry methods to verify the availability, completeness, and accuracy of the eSource technology.

In the network environment of this project, the technology provider accessed the hospital network through a virtual private network (VPN) and a bastion host, and processed the data of this project as a private cloud, thus ensuring the security of the hospital data.

Data Integration

The hospital information system involved in this project has reached the national standards of "Level 3 Equivalence," "Electronic Medical Record Level 5," and "Interoperability Level 4." The medical data intelligence platform in this project is deployed in a hospital intranet, isolated from external networks. Integrated data from different information systems, including the hospital information system, LIMS, picture archiving and communication system, etc, were deidentified from the platform and transferred to a third-party private cloud platform for translation and data format conversion after authorization by the hospital through a VPN.

The scope of data collection in this project was limited to patients who signed Informed Consent Form (Version 4) for study B7461024. The structured data of four CRF data modules (demographic, concomitant medications, local lab, and vital signs) were extracted from the source data in hospital systems, and data processing was completed.

Three-Layer Deidentification of Data

In this project, three layers of deidentification were performed on the electronic source data to ensure data security. The first layer of deidentification was performed before the certified copy of data was loaded to the hospital's medical data intelligence platform. The second layer of deidentification follows the Health Insurance Portability and Accountability Act (HIPAA) by deidentifying 18 data fields at the system level. A third layer of deidentification was performed when mapping and transforming third-party databases for the clinical trial data (demographics, concomitant medications, laboratory tests, and vital signs) collected for this study, as required by the project design.

Collected data did not contain any sensitive information with personal identifiers of the patients, and all deidentification processes were conducted in the internal environment of the hospital. In addition to complying with the relevant laws and regulations, we followed the requirements of Good Clinical Practice regarding patient privacy and confidentiality, and further complied with the requirements of HIPAA to deidentify the 18 basic data fields. Data fields outside the scope of HIPAA will be deidentified and processed in accordance with the TransCelerate guidelines published in April 2015 to ensure the security of patients' personal information and to eliminate the possibility of patient information leakage [8].

The general rules for the third layer of deidentification were as follows:

- Time field: A specific time point is used as the base time, and the encrypted time value is the difference between the word time and the base time
- ID field: Categorized according to the value and only shows the category
- Age field: Categorized according to the value and only shows the category
- Low-frequency field: set to null

In addition, all data flows keep audit trails throughout and are available for audit.

Data Normalization and Information Extraction

After three layers of deidentification, the data was transferred from a hospital to a third-party private cloud platform through a VPN, where translation from Chinese to English and data format conversion were implemented. The whole transfer process was performed for the data that was collected for the clinical trial of this study. Standardization of data is a crucial task during the data preparation phase. This process involves consolidating data from different systems and structures into a consistent, comprehensible, and operable format. First, a thorough examination of data from various systems is necessary. Understanding the data structure, format, and meaning of each system is essential. The second step involves establishing a data dictionary that clearly outlines the meaning, format, and possible values of each data element. Next, selecting a data standard is necessary to ensure consistency and comparability. In this study, we adopted the Health Level 7 (HL7) standard. Additionally, data cleansing and transformation are needed to meet standard requirements, including handling missing data, resolving mismatched data formats, or performing data type conversions. Extract, transform, and load tools were used to integrate data from different systems. Data security must be ensured throughout the data integration process. This includes encrypting sensitive information and strictly managing data permissions. Data verification and validation steps were then performed by professional staff on the translated data. The data from the hospital's medical data intelligence platform were then converted from JSON format to XML and Excel formats. The processed data was transferred back to the hospital via a VPN to a designated location for final adjudication before loading to the sponsor's environment.

One-Time Data Push and Quality Assessment

After the hospital received the processed data, it was then pushed by the hospital to the sponsor's secure and controlled environment (Figure 1). All data deidentification processes were conducted in the hospital's environment, and none of the data obtained by the sponsor can be traced back to patients' personal information to ensure their privacy and information security.

The data quality of this project was assessed using industry data quality assessment rules [9], which are shown in Table 1.

Figure 1. Project operation flow. EHR: electronic health record.

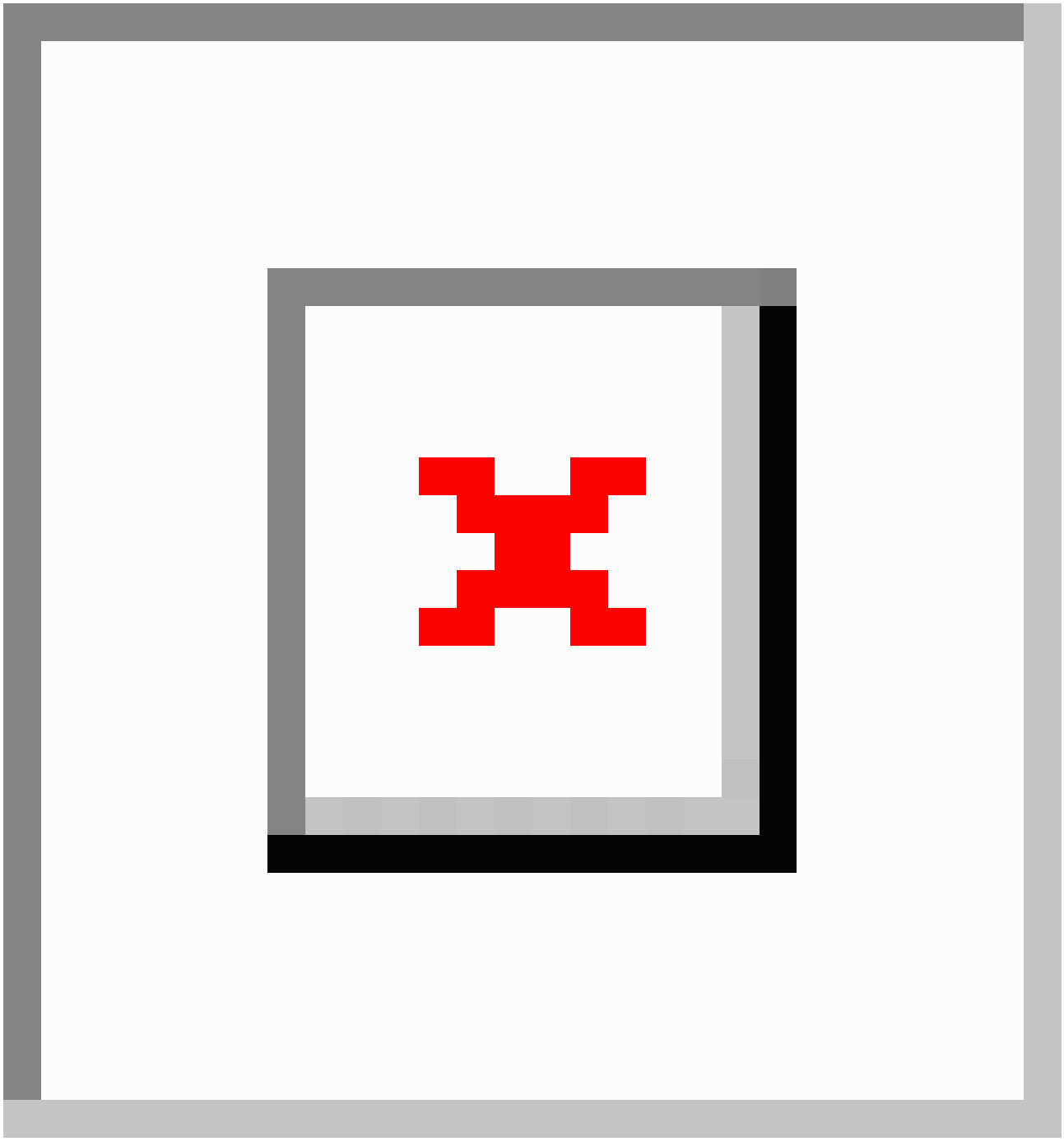


Table . Introduction of data quality assessment rules.

Data validation methods	Dimension	Method description	Cases
Data availability verification	Field dimension	The ratio of the total number of data fields in the clinical trial CRF[a] available in the hospital EHR[b] to the total number of data fields required in the electronic CRF: EHR[c]/CRF[d] × 100%	Based on the electronic CRF, 6 data fields in the demography need to be captured, and 3 of them have records in the EHR. Data availability: 3/6 × 100% = 50%
Data availability verification	Field dimension	The ratio of the total number of data fields in the clinical trial CRF (eSource) that can be transmitted electronically in the hospital's EHR to the total number of data fields required in the electronic CRF: eSource[e]/CRF[d] × 100%	Based on the electronic CRF, 6 data fields in the demography need to be captured, and 2 data fields can be captured by the eSource method. Data availability: 2/6 × 100% = 33.33%
Data completeness verification	Numerical dimension	The ratio of the total number of nonnull data (eSourceV) captured (processed and sent to the sponsor) via the eSource method to the total number of data fields requested on the electronic CRF: eSourceV[f]/CRF[d] × 100%	Based on the clinical trial design, 38 concomitant medication pages need to be collected: 7 pages were collected via eSource and 2 fields was entered per page. Data completeness: 7 × 2/(2 × 38) × 100% = 18.42%
Data accuracy verification	Numerical dimension	Matching of data field values in the hospital's EHR with data field values that can be captured by eSource (data fields that are processed and sent to the sponsor)	4 fields of demography were successfully transmitted through eSource, with 4 data points in each. After comparing with the data in the electronic data capture system, there were no mismatches for one data point. Data accuracy: 8/(2 × 4) × 100% = 100%

[a]CRF: case report form.

[b]EHR: electronic health record.

[c]Total number of data fields in the hospital's EHR.

[d]Total number of data fields requested in the electronic CRF.

[e]Total number of data fields captured (processed and sent to the sponsor) through the eSource method.

[f]Total number of nonempty data fields captured (processed and sent to the sponsor) through the eSource method.

Results

In this project, we collected patients' demographics, vital signs information, local laboratory data, and concomitant medication data from EHRs, successfully pushed the data directly to the designated sponsor environment, and evaluated the data quality from three perspectives including availability, completeness, and accuracy (Table 2).

- The eSource-CRF availability score, which is used to evaluate the ratio of fields in EHR that can be collected by eSource and used for CRF, was low for demographics, blood tests, and urine sample tests but higher for vital signs and concomitant medications.
- Data completeness, defined as the ratio of the total number of nonnull data captured by eSource to the total number of data fields required in the electronic CRF, was used to evaluate the ratio of nonnull data fields in the CRF that can be captured by eSource. In this study, the completeness score of the vital signs module was only 1.32%, and the concomitant medications and laboratory test modules also had poor performance in the data completeness evaluation.
- Data accuracy, defined as the compatibility between the data field values in the hospital EHR and the data field values that can be collected using eSource, was 100% for all modules.
- EHR-CRF availability, which is used to evaluate the ratio of fields in the EHR that can be used for the CRF, was 50%, 60%, and 66.67% for demographics, blood tests, and urine sample tests, respectively, in this study, and the rest of the data were 100% available.

Table . Metrics measured.

CRF[a] domain		CRF-EHR[b] data availability, n/N (%)	CRF-eSource data availability, n/N (%)	Data completeness (preliminary findings), n/N (%)	Data accuracy (preliminary findings), n/N (%)
Definition		Study CRF data elements available in hospital EHR	Study CRF data elements available in hospital EHR and able to be electronically transferred through eSource technology	Study CRF data elements available and entered into hospital EHR and transferred through eSource technology	Study CRF data elements available and entered into hospital EHR and transferred through eSource technology with expected result (eg, matches what was entered directly in form)
Demographics		3/6 (50.00)	2/6 (33.33)	12/12 (100.00)	12/12 (100.00)
Vital signs		10/10 (100.00)	9/10 (90.00)	24/1812 (1.32)[c]	20/20 (100.00)
Local lab					
	Blood biochemical tests	6/10 (60.00)	5/10 (50.00)	12,968/13,540 (95.78)[d]	7767/7767 (100.00)
	Urine sample tests	6/9 (66.67)	5/9 (55.56)	15/40 (37.56)	15/15 (100.00)
Concomitant medication		10/10 (100.00)	9/10 (90.00)	14/76 (18.42)[e]	6/6 (100.00)

[a]CRF: case report form.

[b]EHR: electronic health record.

[c]Checks were made with the relevant clinical research associates (CRAs) regarding the original data collection and CRF completion methods for the following reasons: vital signs were obtained using paper tracking forms provided by the sponsor as the original data source, and the data may not be transcribed into the hospital information system (HIS) by the researcher. Therefore, data from many visits are not available in the HIS.

[d]A total of 2708 blood biochemistry tests were involved.

[e]Concomitant medication uses tracking forms to record adverse event and ConMed (a paper source), and data may not be transcribed into the HIS. As confirmed by the CRA, the percentage of paper ConMed sources was approximately 80%.

Discussion

Although EHRs have been widely used, the degree of structure of EHR data varies substantially among different data modules. In EHRs, demographics, vital signs, local lab data, and concomitant medications are more structured than patient history or progress notes and often contain unstructured text [10]. Therefore, we selected these 4 well-structured data modules for exploration in this project.

For demographics data, among the 6 required fields (subject ID, date of birth, sex, ethnicity, race, and age), subject ID (subject code number/identifier in the trial, not the patient code number/identifier in the EHR system), ethnicity, and race were not available in the EHR, so the EHR-CRF availability score was 50%. Since this was an exploratory project, the date of birth field was also deidentified and thus could not be collected based on our deidentification rule, so the eSource-CRF availability score was 33%. In the future, the availability score can reach close to 100% by bidirectional design of the EHR and CRF under the premise of obtaining compliance for industrial-level applications.

The low availability score of local laboratory data on EHR-CRFs is due to the lack of required fields in the hospital system; “Lab ID” and “Not Done” do not exist in the LIMS, and for the “Clinically Significant” field, the meaning of laboratory test results needs to be manually interpreted by an investigator, so they cannot be transcribed directly. The availability score of eSource-CRFs was further decreased because the field “Laboratory Name and Address” is not an independent structured field in the EHR. The completeness score of urine sample test data was only 37.56% because during the actual clinical trial, especially amid the COVID-19 pandemic period, patients completed study-related laboratory tests at other sites, and those test results were collected via paper-based reports, so the complete data sets cannot be extracted from the site’s system.

To improve data availability in future applications, clinical trial–specific fields need to be added to EHR designs for those data that require an investigator’s interpretation such as “Clinically Significant,” and data transfer and mapping processes for the determination of the scope of data collection also needs to be optimized. Based on these two conditions, the completeness score can be improved to over 90%.

The availability and accuracy of vital signs data are ideal. However, since not all vital signs data collection was recorded by the electronic system during the actual study visit, many vital signs data were collected in “patient diary” and other types of paper-based documents during the study, resulting in a serious limitation in data completeness. With the development of more clinical trial–related electronic hardware and enhancements in products intelligence, more vital signs data will be directly collected by electronic systems, and the completeness of vital signs data transferred from EHR to EDC will be greatly improved in the future.

In the concomitant medication module, there was a good score for availability and accuracy because the standardization and structuring of prescriptions are well done in this hospital system. However, the patient's medication use period during hospitalization is recorded in unstructured text, so the data could not be captured for this study, resulting in a low completeness score of 18.42% for concomitant medication.

In summary, the accuracy score of eSource data in this study was high (100% for all fields). A study by Memorial Sloan Kettering Cancer Center and Yale University confirmed that the error rate of automatic transcription reduced from 6.7% to 0% compared to manual transcription [10]. However, data availability and completeness have not reached a good level. Data availability varies widely across studies, ranging from 13.4% in the Retrieving EHR Useful Data for Secondary Exploitation (REUSE) project [11] to 75% in The STARBRITE Proof-of-Concept Study [12], mainly related to the coverage and structure of the EHR.

National drug regulatory agencies (eg, US Food and Drug Administration [FDA], European Medicines Agency, Medicines and Healthcare products Regulatory Agency, and Pharmaceuticals and Medical Devices Agency) have developed guidelines to support the application of eSource to clinical trials [3,13-15]. The new Good Clinical Practice issued by the Center for Drug Evaluation in 2020 encourages investigators to use clinical trials' electronic medical records for source data documentation [1]. Despite this, we still encountered challenges, including ethical review and data security, during this study's implementation process. Without knowing the precedents, the project team decided to follow the requirements for clinical trials to control the quality of the study. There were no existing regulatory policies or national guidance on eSource in China at the time of this study. The project team provided explanations for inapplicable documents and communicated several times to ensure the approval of relevant institutional departments before finally becoming the first eSource technology study to be approved by the Ethics Committee and Human Genetic Resource Administration of China.

In the absence of regulatory guidelines, our eSource study, the first in China's International Multi-center Clinical Trial, navigated challenges in data deidentification. We adopted HIPAA and TransCelerate's guidelines [8]. Securing approval under "China International Cooperative Scientific Research Approval for Human Genetic Resources," we answered queries and achieved unprecedented recognition. For transferring data from the hospital to the sponsor's environment, we prioritized security and obtained necessary approvals. Iterative revisions ensured a robust data flow design. Challenges in mapping hospital EHR to EDC standards highlighted the need for a scalable mechanism. This study pioneers eSource tech integration in China, emphasizing the importance of seamless data mapping. In the process of executing data standardization, several challenges may arise, including inconsistent data definitions. Data from different systems may use different definitions due to the independent development of these systems, leading to varied interpretations of even identical concepts. To address this issue, establishing a unified data dictionary is crucial to ensure consensus on the definition of each data element. Different systems might also use distinct data formats such as text encodings. Preintegration format conversion is required, and extract, transform, and load tools or scripts can assist in standardizing these formats. During the integration of data from multiple systems, it is possible to discover data in one system that is not present in another. In the data standardization process, considerations must be made on how to handle missing data, which may involve interpolation, setting default values, etc. Quality issues like errors, duplicates, or inaccuracies may exist in data from different systems. Data cleansing, involving deduplication, error correction, logical validation, etc, is necessary to address these quality issues. Different systems may generate data based on diverse business rules and hospital use scenarios. In data standardization, unifying these rules requires collaboration with domain experts to ensure consistency.

Internationally, multiple research studies and publications have been released on regulations, guidelines, and validation of eSource. The FDA provided guidance on the use of electronic source data in clinical trials in 2013 that aims to address barriers to capturing electronic source data for clinical trials, including the lack of interoperability between EHRs and EDC systems. The European-wide Electronic Health Records for Clinical Research (EHR4CR) project was launched in 2011 to explore technical options for the direct capture of EHR data within 35 institutions, and the project was completed in 2016 [16]. The second phase of the project connected the EHRs to EDC systems [17] and aimed to realize the interoperability of EHRs and EDC systems. The US experience focuses more on improving and standardizing the existing EHRs to make them more uniform.

In Europe, the experience focuses on breaking down the technical barrier of interoperability between EHRs and EDC systems. In China, the current industry trends focus on the governance of existing EHR data in the hospital and the building of clinical data repository platforms [7]. Clinical data repository platforms focus on data integration and cleaning between EHRs and other systems in hospital environments and on unstructured data normalization and standardization by natural language processing and other AI technology [18]. At the national level, China is also actively promoting the digitization of medical big data and is committed to the formation of regional health care databases [19], which lays the foundation for the future implementation of eSource in China [20].

This study evaluates the practical application value of eSource in terms of availability, completeness, and accuracy. To improve availability, the structure of the CRF needs to be designed according to the information of the EHR data at the design stage of clinical trials. Even so, since EHRs are designed for the physicians to conduct daily health care activities, certain fields in clinical trials (eg, judgment of normal or abnormal values of laboratory tests and judgment of correlations of adverse events and combined medications) are still not available, and clinical trial–specific fields need to be added to EHR designs for those data that require investigators' interpretation to improve data availability. Completeness could be improved by the development of hospital digitalization that ensures patients' data is collected electronically rather than on paper. Additionally, 2708 blood test records were successfully collected from only 6 patients via eSource in this study, which indicates

that laboratory tests often contain large amounts of highly structured data that are suitable for eSource. EHR-EDC end-to-end automatic data extraction by eSource is suitable for laboratory examinations and can improve the efficiency and accuracy of data extraction significantly as well as reduce redundant manual transcriptions and labor costs. Processing unstructured or even paper-based data in eSource is still a big challenge. Using machine learning tools (eg, natural language processing tools) for autostructuring can be explored in the future. The goal is to have common data standards and better top-level design to facilitate data integrity, interoperability, data security, and patient privacy protection in eSource applications. During deidentification, we processed certain data with a specific logic to protect privacy. The accuracy assessment was performed during the deidentification step to ensure that the data was still sufficiently accurate while meeting privacy requirements. Reversible methods need to be used when performing deidentification as well as providing controlled access mechanisms to the data so that the raw data can be accessed when needed. It is worth noting that different regions and industries may have different privacy regulations and compliance requirements. When deidentifying, you need to ensure that you are compliant with the relevant regulations and understand the limitations of data use. This may require working closely with a legal team.

In the future, we can consider adding performance analysis, including an assessment of data import performance. This involves evaluating the speed and efficiency of data import to ensure it is completed within a reasonable timeframe. Additionally, analyzing data query performance is crucial in practical applications to ensure that the imported data meets the expected query performance in the application. For long-term applications involving a larger size of patients, it is advisable to consider adding analyses related to maintainability and cost-effectiveness. This includes implementing detailed logging and monitoring mechanisms to promptly identify and address potential issues. Furthermore, for the imported data, establishing a version control mechanism is essential for tracing and tracking changes in the data. Simultaneously, for overall resource use, evaluating the resources required during the data import process ensures completion within a cost-effective framework. It is also important to consider the value of imported data for clinical trial operations and related decision-making, providing a comparative analysis between cost and value.

Acknowledgments

This research was supported by the Capital's Funds for Health Improvement and Research (grant No. CFH2022-2Z-2153), and the Beijing Municipal Science & Technology Commission (grant No. Z211100003521008).

Conflicts of Interest

None declared.

References

1. Good clinical practice for drug clinical trial (GCP) 2020[EB/OL]. National Medical Products Administration. 2020 Apr 26. URL: https://www.nmpa.gov.cn/xxgk/fgwj/xzhgfxwj/20200426162401243.html [accessed 2024-06-07]
2. Sheng Q, Wang B, Chen J, et al. Classification and application of electronic source data in clinical trials [Article in Chinese]. Chin Food Drug Admin Magazine 2021;3:36-43 [FREE Full text]
3. Guidance for industry: electronic source data in clinical investigations. Food and Drug Administration. 2013 Sep. URL: https://www.fda.gov/media/85183/download [accessed 2024-06-07]
4. Safran C, Bloomrosen M, Hammond WE, et al. 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. [doi: 10.1197/jamia.M2273] [Medline: 17077452]
5. Garza M, Myneni S, Fenton SH, Zozus MN. eSource for standardized health information exchange in clinical research: a systematic review of progress in the last year. J Soc Clin Data Manag 2021;1(2). [doi: 10.47912/jscdm.66]
6. Beresniak A, Schmidt A, Proeve J, et al. Cost-benefit assessment of using electronic health records data for clinical research versus current practices: contribution of the electronic health records for clinical research (EHR4CR) European project. Contemp Clin Trials 2016 Jan;46:85-91. [doi: 10.1016/j.cct.2015.11.011] [Medline: 26600286]
7. Dong C, Yao C, Gao S, et al. Strengthening clinical research source data management in hospitals to promote data quality of clinical research in China. Chin J Evid Based Med 2019;19:11-1261 [FREE Full text]
8. Data de-identification and anonymization of individual patient data in clinical studies: a model approach. TransCelerate BioPharma. 2015. URL: http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-Data-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies-1.pdf [accessed 2024-06-13]
9. Nordo A, Eisenstein EL, Garza M, Hammond WE, Zozus MN. Evaluative outcomes in direct extraction and use of EHR data in clinical trials. Stud Health Technol Inform 2019;257:333-340. [Medline: 30741219]
10. Vattikola A, Dai H, Buckley M, Maniar R. Direct data extraction and exchange of local LABS for clinical research protocols: a partnership with sites, biopharmaceutical firms, and clinical research organizations. J Soc Clin Data Manag 2021 Mar;1(1). [doi: 10.47912/jscdm.21]

11. El Fadly A, Rance B, Lucas N, et al. Integrating clinical research with the healthcare enterprise: from the RE-USE project to the EHR4CR platform. J Biomed Inform 2011 Dec;44 Suppl 1:S94-S102. [doi: 10.1016/j.jbi.2011.07.007] [Medline: 21888989]
12. Kush R, Alschuler L, Ruggeri R, et al. Implementing single source: the STARBRITE proof-of-concept study. J Am Med Inform Assoc 2007;14(5):662-673. [doi: 10.1197/jamia.M2157] [Medline: 17600107]
13. Reflection paper on expectations for electronic source data and data transcribed to electronic data collection tools in clinical trials. European Medicines Agency. 2010 Jun 9. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/reflection-paper-expectations-electronic-source-data-data-transcribed-electronic-data-collection_en.pdf [accessed 2024-06-07]
14. Technical conformance guide on electronic study data submissions. Pharmaceuticals and Medical Devices Agency. 2015 Apr 27. URL: https://www.pmda.go.jp/files/000206449.pdf [accessed 2024-06-07]
15. MHRA position statement and guidance: electronic health records. MHRA. 2015 Sep 16. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/470228/Electronic_Health_Records_MHRA_Position_Statement.pdf [accessed 2024-06-07]
16. McCowan C, Thomson E, Szmigielski CA, et al. Using electronic health records to support clinical trials: a report on stakeholder engagement for EHR4CR. Biomed Res Int 2015 Oct;2015:707891. [doi: 10.1155/2015/707891] [Medline: 26539523]
17. Sundgren M, Ammour N, Hydes D, Kalra D, Yeatman R. Innovations in data capture transforming trial delivery. Appl Clin Trials 2021 Aug 12;30(7/8) [FREE Full text]
18. Wang Q, Yingping Y. Governance and application of big data in clinical healthcare. J Med Inform 2018;39:2-6. [doi: 10.3969/j.issn.1673-6036.2018.08.001]
19. Guidance from the general office of the state council on promoting and regulating the development of health care big data applications 2016[EB/OL]. Gov.CN. 2016. URL: https://www.gov.cn/gongbao/content/2016/content_5088769.htm [accessed 2024-06-07]
20. Wang B, Lai J, Liao X, Jin F, Yao C. Challenges and solutions in implementing eSource technology for real-world studies in China: qualitative study among different stakeholders. JMIR Form Res 2023 Aug 10;7:e48363. [doi: 10.2196/48363] [Medline: 37561551]

Abbreviations

ALK: anaplastic lymphoma kinase
CRC: clinical research coordinator
CRF: case report form
EDC: electronic data capture
EHR: electronic health record
EHR4CR: Electronic Health Records for Clinical Research
FDA: Food and Drug Administration
HIPAA: Health Insurance Portability and Accountability Act
HL7: Health Level 7
LIMS: laboratory information management system
REUSE: Retrieving EHR Useful Data for Secondary Exploitation
VPN: virtual private network

Edited by C Lovis; submitted 19.09.23; peer-reviewed by H Veldandi, Y Su; revised version received 20.12.23; accepted 18.04.24; published 27.06.24.

Please cite as:
Yuan Y, Mei Y, Zhao S, Dai S, Liu X, Sun X, Fu Z, Zhou L, Ai J, Ma L, Jiang M
Data Flow Construction and Quality Evaluation of Electronic Source Data in Clinical Trials: Pilot Study Based on Hospital Electronic Medical Records in China
JMIR Med Inform 2024;12:e52934
URL: https://medinform.jmir.org/2024/1/e52934
doi:10.2196/52934

Examining Linguistic Differences in Electronic Health Records for Diverse Patients With Diabetes: Natural Language Processing Analysis

Isabel Bilotta[1], PhD; Scott Tonidandel[2], PhD; Winston R Liaw[3], MPH, MD; Eden King[4], PhD; Diana N Carvajal[5], MPH, MD; Ayana Taylor[6], MD; Julie Thamby[7], BA; Yang Xiang[8], PhD; Cui Tao[9], PhD; Michael Hansen[10], MPH, MS, MD

1
2
3
4
5
6
7
8
9
10

Corresponding Author:
Winston R Liaw, MPH, MD

Abstract

Background: Individuals from minoritized racial and ethnic backgrounds experience pernicious and pervasive health disparities that have emerged, in part, from clinician bias.

Objective: We used a natural language processing approach to examine whether linguistic markers in electronic health record (EHR) notes differ based on the race and ethnicity of the patient. To validate this methodological approach, we also assessed the extent to which clinicians perceive linguistic markers to be indicative of bias.

Methods: In this cross-sectional study, we extracted EHR notes for patients who were aged 18 years or older; had more than 5 years of diabetes diagnosis codes; and received care between 2006 and 2014 from family physicians, general internists, or endocrinologists practicing in an urban, academic network of clinics. The race and ethnicity of patients were defined as *White non-Hispanic*, *Black non-Hispanic*, or *Hispanic or Latino*. We hypothesized that Sentiment Analysis and Social Cognition Engine (SEANCE) components (ie, negative adjectives, positive adjectives, joy words, fear and disgust words, politics words, respect words, trust verbs, and well-being words) and mean word count would be indicators of bias if racial differences emerged. We performed linear mixed effects analyses to examine the relationship between the outcomes of interest (the SEANCE components and word count) and patient race and ethnicity, controlling for patient age. To validate this approach, we asked clinicians to indicate the extent to which they thought variation in the use of SEANCE language domains for different racial and ethnic groups was reflective of bias in EHR notes.

Results: We examined EHR notes (n=12,905) of Black non-Hispanic, White non-Hispanic, and Hispanic or Latino patients (n=1562), who were seen by 281 physicians. A total of 27 clinicians participated in the validation study. In terms of bias, participants rated negative adjectives as 8.63 (SD 2.06), fear and disgust words as 8.11 (SD 2.15), and positive adjectives as 7.93 (SD 2.46) on a scale of 1 to 10, with 10 being extremely indicative of bias. Notes for Black non-Hispanic patients contained significantly more negative adjectives (coefficient 0.07, SE 0.02) and significantly more fear and disgust words (coefficient 0.007, SE 0.002) than those for White non-Hispanic patients. The notes for Hispanic or Latino patients included significantly fewer positive adjectives (coefficient −0.02, SE 0.007), trust verbs (coefficient −0.009, SE 0.004), and joy words (coefficient −0.03, SE 0.01) than those for White non-Hispanic patients.

Conclusions: This approach may enable physicians and researchers to identify and mitigate bias in medical interactions, with the goal of reducing health disparities stemming from bias.

(JMIR Med Inform 2024;12:e50428) doi:10.2196/50428

KEYWORDS

bias; sociodemographic factors; health care disparities; natural language processing; sentiment analysis; diabetes; electronic health record; racial; ethnic; diversity; Hispanic; medical interaction

Introduction

Background

Language and communication play a significant, if not primary, role in social relations across different cultures [1]. Language has increasingly been recognized as a relevant form of data that describe relations and behavior [2]. One of the most intimate forms of communication between individuals occurs between clinicians and patients during clinical visits. However, these encounters may be undermined by different forms of bias directed toward patients from certain racial and ethnic minority groups [3]. Generally, *bias* refers to an evaluation, decision, perception, or action in favor of or against a person or group compared to another. Bias can be blatant, wherein it is characterized by deliberate actions (eg, racist comments) that are intentionally and overtly discriminatory [4]. Bias can also be subtle, including "actions that are ambiguous in intent to harm, difficult to detect, low in intensity, and often unintentional but are nevertheless deleterious" to targets [4]. Subtle bias by health care clinicians is linked to negative outcomes for racial and ethnic minority patients, particularly Black non-Hispanic and Hispanic or Latino patients [5].

Race and Racial Bias in Medical Interactions

Health disparities between racial and ethnic groups have historically been attributed to varying levels of socioeconomic status, as well as genetic and biological factors that were thought to predispose groups to different medical conditions. Research has emerged over the past few decades demonstrating that in fact, there is no biological basis for racial and ethnic differences. Humans share 99.9% of their genome, and the 0.1% variation cannot be explained or elucidated by race [6]. Race describes physical traits considered socially significant, and ethnicity denotes a shared cultural heritage, such as language, practices, and beliefs [7]. As such, race and ethnicity are social constructs, and since the landmark report *Unequal Treatment* in 2002 detailed the impact of racial and ethnic discrimination in patient-clinician interactions, research interest in this area has burgeoned [8]. Relative to White non-Hispanic patients, Black non-Hispanic and Hispanic or Latino patients are less likely to ''engender empathic responses from clinicians, establish rapport with clinicians, receive sufficient information, and be encouraged to participate in medical decision making" [9]. A lack of relationship building [10], reduced positive patient and clinician affect [11], decreased patient trust [12], and fewer patient questions [13] are all more likely outcomes for Black non-Hispanic and Hispanic or Latino patients compared to White non-Hispanic patients during medical interactions. Indeed, the 2018 *National Healthcare Disparities Report* revealed that, compared to White non-Hispanic patients, Black non-Hispanic patients receive inferior care on 40% of quality measures, and Hispanic or Latino patients receive worse care on 35% of quality measures, many of which indicate biased and discriminatory behaviors by clinicians [14]. For example, indicators were worse for Black non-Hispanic and Hispanic or Latino patients than White non-Hispanic patients for measures such as "physicians sometimes or never showed respect for what they had to say" and "physicians sometimes or never spent enough time with them" [14]. Black non-Hispanic and Hispanic or Latino patients are more likely to report racial and ethnic bias and discrimination during medical encounters compared to White non-Hispanic patients [15]. Yet, less is known about the manifestations and details of such experiences during the clinician-patient interaction [16] and whether racial and ethnic discrepancies in care can be observed in the content of electronic health records (EHRs). Similar to the thesis described in *Unequal Treatment*, we hypothesized that the mitigation of bias at the clinician level is needed to improve patient outcomes for diverse racial and ethnic populations and narrow the disparities gap. To address bias, researchers need to understand how to measure its existence, and clinicians need to be informed of its manifestations.

Research Contributions

Bias can have many forms—blatant, subtle, malevolent, or benevolent—all of which can be indicated by language. With increasing access to EHR documentation and advances in natural language processing, we may be better equipped to identify differences in clinician encounters with patients of diverse racial and ethnic backgrounds. This study searched for linguistic discrepancies in EHRs using a natural language processing approach followed by linear mixed effect model analyses. EHRs are digital summaries of the clinician-patient encounter and include the clinician's assessment of the interaction, as well as the patient's health history. Since the clinician is responsible for inputting information, as well as reviewing the information inputted by other care clinicians in the EHR for each patient encounter, the contents of the EHR may be particularly useful in illuminating biases that clinicians hold toward patients of different racial and ethnic backgrounds. Although several studies have indicated that clinician bias occurs, particularly in racially and ethnically discordant interactions (ie, when the patient and clinician are of different racial and ethnic backgrounds), relatively little research has examined the ways in which the clinician may be thinking about the patient and how the clinician's sentiment and cognitions are reflected in the language of the EHR [8,17]. EHRs can include many years of patient-clinician interactions, with multiple clinicians having access to them, allowing for biases to be passed on and potentially impact future medical decisions.

Our data set contained EHR notes for a large sample of White non-Hispanic, Black non-Hispanic, and Hispanic or Latino patients with diabetes in the Southern United States. The natural language processing tool, Sentiment Analysis and Social Cognition Engine (SEANCE), was applied to assess multiple linguistic markers in the EHR text [18,19]. We then explored whether 8 of the 20 SEANCE components (see Table 1) differed for patients of different races and ethnicities [20,21].

Table . Description of SEANCE[a] components.

Component label	Indices, n[b]	Key indices[c]	Language examples
Negative adjectives	18	NRC[d] negative adjectives, NRC disgust adjectives, NRC anger adjectives, GI[e] negative adjectives, and Hu-Liu[f] negative adjectives	Unkind, bad, cruel, hurtful, and intolerant
Positive adjectives	9	Hu-Liu positive adjectives, VADER[g] positive adjectives, GI positive adjectives, and Lasswell[h] positive affect adjectives	Supportive, kind, great, and nice
Joy words	8	NRC joy adjectives, NRC anticipation adjectives, and NRC surprise adjectives	Admiration, advocacy, elated, glad, liking, and pleased
Fear and disgust words	8	NRC disgust nouns, NRC negative nouns, NRC fear nouns, and NRC anger nouns	Abnormal, adverse, attack, cringe, criticize, distress, intimidate, unequal, and stigma
Politics words	7	GI politics nouns and Lasswell power nouns	Alliance, ally, authorize, civil, concession, consent, and oppose
Respect words	4	Lasswell respect nouns	Status, honor, recognition, and prestige
Trust verbs	5	NRC trust verbs, NRC joy verbs, and NRC positive verbs	Affirm, advise, confide, and cooperating
Well-being words	4	Lasswell well-being physical nouns and Lasswell well-being total nouns	Alive, ambulance, adjust, afraid, blood, clinic, and nutrition

[a]SEANCE: Sentiment Analysis and Social Cognition Engine.

[b]Indices refer to the number of dictionary lists from which the component was developed.

[c]The key indices came from the following dictionary lists: NRC Emotion Lexicon [18,22], the Harvard-IV dictionary list used by the General Inquirer [23], the Hu-Liu polarity word lists [22,23], the Valence Aware Dictionary and Sentiment Reasoner [24], the Lasswell dictionary lists [25,26], and the Geneva Affect Label Coder database [27]. For a thorough review of the SEANCE indices and corresponding dictionaries, see Crossley et al [18].

[d]NRC: NRC Emotion Lexicon.

[e]GI: General Inquirer.

[f]Hu-Liu: Hu-Liu polarity word lists.

[g]VADER: Valence Aware Dictionary and Sentiment Reasoner.

[h]Lasswell: Lasswell dictionary lists.

We hypothesized that the SEANCE components for negative adjectives, positive adjectives, joy words, fear and disgust words, politics words, respect words, trust verbs, and well-being words and the mean word count in the notes would be indicators of bias, as these concepts have been linked to bias in nonmedical contexts. Ng's [28] review of linguistic racial bias in verbiage offers the rationale for our choice of fear and disgust words, politics words, respect words, and trust verbs as indicators of bias, whereas the work of Li et al [29] examining gender differences in standardized writing assessment provides further support for our use of SEANCE as a tool for examining biases in language. We selected positive and negative adjectives, well-being words, politics words, and word count indicators as prior research demonstrates that clinicians may be less likely to establish rapport and provide appropriate medications and are more inclined to show negative attitudes and be dismissive toward Black non-Hispanic and Hispanic or Latino patients as a result of their unconscious racial and ethnic biases [30-33].

Specifically, we investigated which aspects of communication differ and whether differences are indicative of biased interactions. Any systematic variation in language can convey differential perceptions, attitudes, and expectations. For example, words such as "resistant" or "non-compliant" could reflect bias if (all else being equal) they tend to be used more to reflect people from some racial or ethnic backgrounds than others. This work aimed to elucidate for clinicians and researchers where discrepancies in communication emerge in the EHR and whether these differences are indicative of racial and ethnic bias. We also assessed the extent to which clinicians perceive linguistic markers to be indicative of bias.

Methods

Sample

This was a cross-sectional study using EHR-derived physician notation of outpatient clinical encounters. We extracted EHR encounters (n=15,460) for patients (n=1647) who were aged 18 years or older; had more than 5 years of diabetes diagnosis codes; and received care between 2006 and 2014 from family physicians, general internists, or endocrinologists practicing in an urban, academic network of clinics. We chose this disease

because of its high prevalence (11.3% in the United States) and chose to examine outpatient visits because of the relative scope of annual outpatient visits (1 billion) relative to hospital admissions (32 million) [34-36]. The demographic variables collected were patient race and ethnicity, sex, and age. The race and ethnicity of patients were defined as *White non-Hispanic*, *Black non-Hispanic*, or *Hispanic or Latino* (see Table 2 for a summary of patient demographics).

Table . Patient demographics of the final sample.

Variable		Value (n=1562)
Age (years)		
	Mean (SD)	68.74 (13.76)
	Range	20-102
	Median (IQR)	69 (61-78)
Sex, n (%)		
	Female	871 (55.74)
	Male	691 (44.26)
Race and ethnicity, n (%)		
	White non-Hispanic	682 (43.66)
	Black non-Hispanic	755 (48.34)
	Hispanic or Latino	125 (8)

SEANCE Algorithm

SEANCE is a lexical scoring algorithm that includes over 200 word vectors (also referred to as indices or features) designed to assess sentiment, cognition, and social order, which were developed from preexisting and widely used databases such as EmoLex and SenticNet [22,37]. In addition to the core indices, SEANCE allows for several customized indices, including filtering for particular parts of speech and controlling for instances of negation [18]. Since SEANCE computes such a large quantity of indices, Crossley et al [18] developed 20 components from all the indices using principal component analysis (PCA) [18]. These components are essentially clusters of related indices in SEANCE and allow users to interpret the SEANCE output at a more macro level. This process enabled them to summarize the SEANCE indices into a smaller and more interpretable set of variables. In the PCA by Crossley et al [18], they retained even the smallest components, setting a conservative cutoff point for inclusion (ie, 1% for variance explained by each component). The analyses for this research were run on a subset of 8 of the 20 *components* that Crossley et al [18] developed. We selected these 8 components a priori (see Table 1 for a description of the selected components).

We chose SEANCE instead of other natural language processing tools, such as Linguistic Inquiry and Word Count (LIWC), because it contains a larger number of core indices taken from multiple lexicons, as well as 20 components, and is based on the most recent improvements in sentiment analysis [18]. In their validation of SEANCE, Crossley et al [18] found that SEANCE components demonstrated significantly greater accuracy than LIWC indices (P<.001) for 3 of the 4 review types examined. In addition to the core indices, SEANCE allows for several customized indices, including filtering for parts of speech (also known as "parts-of-speech tagging") and controlling for instances of negation, which LIWC does not offer. We analyzed all words in the EHR (ie, *not* single parts of speech), but we controlled for negation. For example, this means that "not good" would be recognized as *not being positive* by SEANCE, as opposed to LIWC, which would see the word "good" and count it as positive.

Validation of the Sentiment Analysis Approach

To provide validation of the sentiment analysis approach used in this study, we surveyed subject-matter experts in EHR note writing (ie, physicians, physician assistants, and nurse practitioners) to garner their perspectives on the appropriateness of the linguistic components identified in our pilot study as indicators of subtle racial and ethnic bias in EHR notes. The team of researchers for this study included industrial-organizational psychologists who have expertise in bias and discrimination; however, it was also valuable to garner opinions from clinicians who are experts in EHR note writing and who understand the differences in the types of language used. To recruit participants, we used a combination of opportunistic and snowball sampling, starting with individuals within our personal networks. Through a web-based program, we asked participants to indicate the extent to which they thought the language domains (eg, negative adjectives, fear and disgust words, etc) were reflective of bias in EHR notes. Participants were told the following:

> *One type of language that could represent bias reflects the amount of NEGATIVE ADJECTIVES contained in the electronic health record. Examples of negative adjectives include "unkind," "bad," "harmful," "intolerant," and "stupid." If these kinds of words were used to describe Black or LatinX patients more than White patients, to what extent do you think this would be indicative of racial bias? Please indicate the extent of your agreement on the 1 to 10 scale below.*

The same formatting was used for each of the linguistic components, with component-specific language examples offered so participants understood the types of sentiment that each component was designed to assess.

Cross-Classified Linear Mixed Effects Models

We used the *lme4* package in R (R Foundation for Statistical Computing) to perform linear mixed effects analyses of the relationships between the outcomes of interest (SEANCE components and word count) and patient race and ethnicity, controlling for patient age. We ran an identical analysis, treating 8 different SEANCE components and the mean word count in the EHR as the dependent variables, while leaving all other variables consistent across the models. The same steps of entering fixed and random effects were applied across all cross-classified linear mixed effects models with different dependent variables (ie, negative adjectives, positive adjectives, well-being words, trust verbs, fear and disgust words, joy words, politics words, respect words, and mean word count).

We first ran a null model with only the random intercepts. We then added random effects and applied a crossed design (vs a traditional nested structure), leading us to have intercepts for physicians and patients. Then, we ran a model with the random intercepts as well as the fixed effects. As fixed effects, we entered *race and ethnicity* and *age* (without an interaction term) into the model. For all models examined, the intercept variation can be attributed primarily to different physicians rather than patients. We used a 95% CI to determine statistical significance. To be more conservative, given that we ran multiple tests, we also computed an additional set of CIs at the 99th percentile.

Ethical Considerations

We obtained ethics approval from the University of Texas Health Science Center's Committee for the Protection of Human Subjects (HSC-MS-18-0431) and the Rice University Institutional Review Board (IRB-FY2021-325). Participants consented and received a US $25 gift card after completing the survey. EHR data were deidentified prior to the analysis.

Results

Description and Justification for Cross-Classified Analyses

An initial inspection of the data revealed that 2 physicians were extreme outliers, accounting for 16.53% (2555/15,460) of the notes in our sample. To ensure that the overrepresentation of these physicians would not bias the results, we removed those notes from the data set (taking us from our initial sample of 15,460 visits with 283 physicians and 1647 patients to 12,905 visits with 281 physicians and 1562 patients; Table 2). The distribution of visits by patients indicates an average of 8.27 visits per patient with a minimum of 1, a median of 5, and a maximum of 97. Physicians see 11.72 patients on average, with a median of 2 and a maximum of 143, suggesting a skewed distribution. Despite the relatively large number of patients seen by some physicians, these physicians accounted for substantially fewer patient notes than the 2 physicians that were previously removed. Patients see 2.11 physicians on average, with a minimum of 1 and a maximum of 12; however, the distribution suggests that 6.6% (109/1647) of patients saw 5 or more physicians. Moreover, 742 (45.1%) of the 1647 patients saw 1 physician, whereas 119 (7.2%) saw 4 physicians. In our data set, patients can have multiple visits to a variety of physicians, indicating that patient visits are not nested within physicians. Further, physicians may see different patients with no consistent overlap of patients between physicians, indicating that physicians are not nested within patients. Thus, there is no clear hierarchical nesting of patients within physicians (or vice versa), which suggests that a cross-classified design is more appropriate than a traditional, hierarchical, multilevel model structure.

Cross-Classified Linear Mixed Effects Model Results

In the negative adjective component model (Table 3), the random effects of patient (σ^2=0.02) and physician (σ^2=0.12) indicated that intercept variation in use of negative adjectives is mainly a function of the physician rather than the patient. The physician random effect was over 5 times as large as the random effect for the patient; the intraclass correlation (ICC) for physicians was 0.41 and the ICC for patients was 0.07 (ICC_{total}=0.481). This pattern of results in random effects and ICC values for patients and physicians was consistent across the other 8 models. Overall, 2 of the 5 relationships (ie, the significant difference in positive adjectives for Hispanic or Latino and White non-Hispanic patient notes, and the significant difference in trust verbs for Hispanic or Latino and White non-Hispanic patient notes) that were previously significant at the 95th percentile had CIs that included zero at the 99th percentile. For 3 of the SEANCE components—well-being, politics, and respect words—and for the overall word count, there was not a statistically significant difference between the 3 races and ethnicities. In contrast, for all the other remaining SEANCE components, there was a statistically significant race and ethnicity effect for either Black non-Hispanic or Hispanic or Latino patients relative to White non-Hispanic patients. Specifically, notes for Black non-Hispanic patients contained significantly more negative adjectives and fear and disgust words than those for White non-Hispanic patients. Notes for Hispanic or Latino patients included significantly fewer positive adjectives, trust verbs, and joy words than those for White non-Hispanic patients. As such, across most of the SEANCE components, we observed favoritism of White non-Hispanic patients in terms of note content.

Table . Fixed effects model results for negative adjectives, positive adjectives, well-being words, trust verbs, joy words, politics words, respect words, fear and disgust words, and word count.

Variables[a]	Negative adjectives	Positive adjectives	Well-being words	Trust verbs	Joy words	Politics words	Respect words	Fear and disgust words	Word count
Fixed effect estimates									
Age (years)									
β (SE)	−0.00 (0.00)	0.00 (0.00)	.0002 (0.00009)	−0.00007 (0.00008)	0.000002 (0.0002)	−0.00009 (0.00004)	−0.00004 (0.00005)	0.000005 (0.00)	−0.43 (0.68)
95% CI	−0.002 to 0.0003	−0.002 to 0.00	0.0006 to 0.0004[b]	−0.002 to 0.0008	−0.0004 to 0.0004	−0.0002 to −0.000007[b]	−0.0004 to 0.0004	−0.0001 to 0.0002	−1.76 to 0.90
Race and ethnicity									
White non-Hispanic (reference)									
β (SE)	0.42 (0.05)	−0.24 (0.017)	0.18 (0.007)	0.16 (0.007)	0.32 (0.02)	0.07 (0.003)	0.05 (0.004)	0.17 (0.007)	868.50 (54.45)
95% CI	0.32 to 0.53	−0.26 to −0.21	0.17 to 0.20	0.14 to 0.17	0.28 to 0.35	0.06 to 0.07	0.04 to 0.05	0.16 to 0.19	761.84 to 975.17
Black non-Hispanic									
β (SE)	0.07 (0.02)	0.02 (0.004)	0.004 (0.002)	−0.003 (0.002)	−0.01 (0.006)	0.001 (0.001)	−0.001 (0.001)	0.007 (0.002)	20.61 (19.01)
95% CI	0.04 to 0.11[b]	−0.006 to 0.01	−0.0007 to 0.009	−0.007 to 0.001	−0.02 to 0.0004	−0.001 to 0.004	−0.004 to 0.002	0.003 to 0.01[b]	−16.71 to 57.84
Hispanic or Latino									
β (SE)	0.02 (0.03)	−0.02 (0.007)	0.002 (0.004)	−0.009 (0.004)	−0.03 (0.01)	−0.0009 (0.003)	0.0006 (0.002)	−0.002 (0.004)	15.73 (32.30)
95% CI	−0.03 to 0.08	−0.03 to −0.004[b]	−0.007 to 0.01	−0.02 to −0.001[b]	−0.05 to −0.01[b]	−0.005 to 0.003	−0.004 to 0.005	−0.01 to 0.006	−47.61 to 78.98
Random effects, estimate (SE)									
U0 patient	0.02 (0.14)	0.0008 (0.03)	0.0004 (0.02)	0.0002 (0.02)	0.0006 (0.02)	0.00001 (0.004)	0.00002 (0.005)	0.0004 (0.02)	27,878 (167.0)
U0 physician	0.12 (0.34)	0.006 (0.08)	0.003 (0.05)	0.003 (0.05)	0.02 (0.15)	0.0002 (0.016)	0.0005 (0.02)	0.003 (0.05)	119,489 (345.7)

[a]Random effects are presented as estimate and SE. For the fixed effect estimates, cell entries are parameter (β) estimates, SE, and 95% CIs. White non-Hispanic was the reference group for race and ethnicity.

[b]Significant effects based on the 95% CIs.

Sentiment Analysis Validation

In all, 27 participants completed the surveys (see Multimedia Appendix 1 for the demographics of the participants). On a scale of 1 to 10, with 10 being extremely indicative of bias, participants rated negative adjectives as 8.63 (SD 2.06), fear and disgust words as 8.11 (SD 2.15), positive adjectives as 7.93 (SD 2.46), trust verbs as 7.56 (SD 2.64), and joy words as 6.81 (SD 2.47). The means and SDs for each of the components are reported in Table 4. The results of this preliminary analysis provide support for the validity of the linguistic components as indicators of bias in EHRs, as our sample of clinicians regard them as highly suggestive of bias if used differently for patients of diverse racial and ethnic backgrounds.

Table . Subject-matter expert assessment of bias based on specific linguistic markers.

Component	Score, mean (SD)[a]
Negative adjectives	8.63 (2.06)
Fear and disgust words	8.11 (2.15)
Positive adjectives	7.93 (2.46)
Joy words	6.81 (2.47)
Trust verbs	7.56 (2.64)
Politics words	7.07 (2.32)
Respect nouns	7.56 (2.55)
Well-being words	5.56 (2.55)
Mean word count	6.11 (2.19)

[a]Scale ranges from 1 (*Not at all indicative of bias*) to 10 (*Extremely indicative of bias*).

Discussion

Principal Findings

We found that the words that physicians use in EHR notes differ based on the racial and ethnic backgrounds of patients. Specifically, for Black non-Hispanic patients, notes consisted of words that convey negativity, fear, and disgust. When seeing Hispanic or Latino patients, physicians used fewer positive words and were less likely to use words that communicate trust and joy. Our findings are consistent with others who have documented that physicians communicate in the EHR differently (ie, more negatively) when caring for patients from some minority groups [9,17], which may ultimately result in adverse and inequitable health outcomes for patients. Our results also align with other papers that found that stigmatizing language is more commonly used in EHRs for minority populations [38-42]. Those papers used language guidelines [38] and experts [39] to identify stigmatizing language. We came to a similar conclusion by using established language dictionaries and contend that our approach allows for a more comprehensive assessment of language. For example, a prior paper used 15 descriptors [42]. In contrast, our approach encompasses tens of thousands of words, including multiple word lists, positive and negative sentiments, and emotions. Thus, this method does not merely capture the presence or absence of stigmatizing language, but rather offers a broader glimpse of the clinician-patient relationship. Furthermore, the validation survey confirmed that subject-matter experts perceive the types of words included in this study to be indicative of bias when used differentially for patients of diverse racial and ethnic backgrounds. Taken together, these findings indicate that the language used differs for patients based on racial and ethnic backgrounds and that those differences are suggestive of bias. As a result, our paper is the first to use this particular method to examine outpatient, diabetes notes. Since diabetes quality measures already exist, our analysis allows researchers to link bias to differences in quality in future studies [43].

EHR notes are important, although imperfect, assessments of physician attitudes toward their patients. With more and more time now being devoted to EHR documentation, physicians are increasingly burned out, which has led to the adoption of more efficient data entry strategies such as using templates, copy-pasting previous text, and inserting preset language [44,45]. Consequently, notes can be standardized, limiting our ability to assess physician attitudes and subconscious biases toward patients. Despite these caveats, notes remain the definitive and often sole account of what happened in the examination room, and based on these data, Black non-Hispanic and Hispanic or Latino patients are written about differently than White non-Hispanic patients.

The method described in this paper offers a scalable blueprint that provides clinicians with data about their interactions with patients and overcomes limitations of other traditional measures of bias. Existing measures require primary data collection through surveys, videotaped encounters, and confederate observations. Surveys assess perceptions of interactions and are prone to retrospective bias and socially desirable responding, whereas the time-consuming nature of encounters and observations lack scalability and limit the number of clinicians that can receive feedback at any given time. The relevance of alternative measures has also been questioned. For example, critics of the implicit association test have asked whether performance on the test is applicable to real-world contexts [46], which may explain why some change their behavior when confronted with their own biases, whereas others do not [5,47]. In contrast, our method uses data that are automatically and universally collected through the course of delivering care and generated by physicians in actual encounters.

Limitations

When interpreting our results, several limitations should be considered. First, due to limitations in our data, we are unable to determine which additional team members, including scribes, medical assistants, and residents, contributed to the notes. However, attending physicians are ultimately responsible for the content and have the authority and responsibility to modify language that is inconsistent with their values. Second, we lack information about physicians in this sample and do not have access to physician demographic characteristics (eg, their racial and ethnic backgrounds), although this would be an important next step. We attempted to account for this limitation by comparing language within rather than across physicians. Third, we included all language within notes, including physical exams,

medications, and past medical histories. These sections can be guided by templates or not actively entered by physicians. We retained these parts in case the language within these sections contributed to variation. An alternative approach could assess only the history of present illness, assessment, and plan sections of the note and could yield different results. Additional work is needed to determine whether differential word choices reflect attitudes and behaviors toward patients. EHR notes serve a wide range of purposes. They convey medical information to others, remind physicians of their impressions, communicate plans to patients, provide justification for billing codes, and serve as legal evidence [44]. Thus, specific phrases (eg, worsening, uncontrolled, or adherence) may be required for billing, compliance, and legal purposes and may not reflect bias toward patients. Finally, these results may not be generalizable to other conditions. Our findings may be unique to the language used for diabetes care and by clinicians who manage diabetes. Determining whether these results persist for different diseases (eg, cancer, heart disease, and acute injuries) is an important next step.

Directions for Future Research

Additional research is needed to interpret and provide context for this exploratory work. To determine whether these measures are associated with bias, subject-matter experts could label notes using known patterns of bias (eg, the ratio of collective to personal pronouns, the amount and level of abstraction of speech, and passive vs active voice) [48]. Further research is needed to understand whether biased language in notes reflects biased behaviors during encounters as well as inequitable health outcomes for some racial and ethnic minority populations. Conducting further experiments (eg, with research actors as patients in a mock medical visit) could help determine whether biased language in notes reflects manifestations of bias during encounters (eg, less eye contact, hostile language, or less time spent on education and counseling). If bias is confirmed, we need to determine whether clinicians who use differential language provide worse care and quality for minority patients. Ultimately, this tool may be used to identify and mitigate bias. Future studies should assess whether receiving feedback using this method leads to behavior change and whether changing the language used in EHR notes leads to changes in patient interactions. Although many strategies for reducing bias exist—such as affirming egalitarian goals, seeking common-group identities, perspective taking, and individuation—it is unclear which approach best complements our proposed method [5].

Conclusion

In this novel, exploratory work, we used natural language processing and found that compared to encounters with White non-Hispanic patients, physicians use language conveying more negativity, fear, and disgust in their encounters with some racial and ethnic minority patients. If confirmed in future studies, these features could be used to make clinicians aware of their biases with the goal of reducing racial and ethnic discrimination and the resulting health inequities.

Acknowledgments

This work was supported by a Rice Anti-Racism Research Grant through Rice University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Demographics of the validation study participants.
[DOCX File, 14 KB - medinform_v12i1e50428_app1.docx]

References

1. Maass A, Karasawa M, Politi F, Suga S. Do verbs and adjectives play different roles in different cultures? a cross-linguistic analysis of person representation. J Pers Soc Psychol 2006 May;90(5):734-750. [doi: 10.1037/0022-3514.90.5.734] [Medline: 16737371]
2. Boroditsky L, Schmidt LA, Phillips W. Sex, syntax, and semantics. In: Gentner D, Goldin-Meadow S, editors. Language in Mind: Advances in the Study of Language and Thought: The MIT Press; 2003. [doi: 10.7551/mitpress/4117.001.0001]
3. Hall WJ, Chapman MV, Lee KM, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. Am J Public Health 2015 Dec;105(12):e60-e76. [doi: 10.2105/AJPH.2015.302903] [Medline: 26469668]
4. Jones KP, Peddie CI, Gilrane VL, King EB, Gray AL. Not so subtle: a meta-analytic investigation of the correlates of subtle and overt discrimination. J Manag 2016 Jul 10;42(6):1588-1613. [doi: 10.1177/0149206313506466]
5. Zestcott CA, Blair IV, Stone J. Examining the presence, consequences, and reduction of implicit bias in health care: a narrative review. Group Process Intergroup Relat 2016 Jul;19(4):528-542. [doi: 10.1177/1368430216642029] [Medline: 27547105]
6. Ahn SM, Kim TH, Lee S, et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 2009 Sep;19(9):1622-1629. [doi: 10.1101/gr.092197.109] [Medline: 19470904]

7. Chadha N, Lim B, Kane M, Rowland B. Toward the abolition of biological race in medicine. Othering & Belonging Institute. 2020 May 13. URL: https://belonging.berkeley.edu/toward-abolition-biological-race-medicine-8 [accessed 2023-06-27]
8. Institute of Medicine. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care: National Academies Press; 2003. [doi: 10.17226/10260]
9. Ferguson WJ, Candib LM. Culture, language, and the doctor-patient relationship. Fam Med 2002 May;34(5):353-361. [Medline: 12038717]
10. Siminoff LA, Graham GC, Gordon NH. Cancer communication patterns and the influence of patient characteristics: disparities in information-giving and affective behaviors. Patient Educ Couns 2006 Sep;62(3):355-360. [doi: 10.1016/j.pec.2006.06.011] [Medline: 16860520]
11. Johnson RL, Roter D, Powe NR, Cooper LA. Patient race/ethnicity and quality of patient–physician communication during medical visits. Am J Public Health 2004 Dec;94(12):2084-2090. [doi: 10.2105/ajph.94.12.2084] [Medline: 15569958]
12. Jacobs EA, Rolle I, Ferrans CE, Whitaker EE, Warnecke RB. Understanding African Americans' views of the trustworthiness of physicians. J Gen Intern Med 2006 Jun;21(6):642-647. [doi: 10.1111/j.1525-1497.2006.00485.x] [Medline: 16808750]
13. Eggly S, Hamel LM, Foster TS, et al. Randomized trial of a question prompt list to increase patient active participation during interactions with Black patients and their oncologists. Patient Educ Couns 2017 May;100(5):818-826. [doi: 10.1016/j.pec.2016.12.026] [Medline: 28073615]
14. National Healthcare Quality & Disparities Reports. Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/research/findings/nhqrdr/index.html [accessed 2023-06-27]
15. Shavers VL, Fagan P, Jones D, et al. The state of research on racial/ethnic discrimination in the receipt of health care. Am J Public Health 2012 May;102(5):953-966. [doi: 10.2105/AJPH.2012.300773] [Medline: 22494002]
16. Penner LA, Dovidio JF, West TV, et al. Aversive racism and medical interactions with Black patients: a field study. J Exp Soc Psychol 2010 Mar 1;46(2):436-440. [doi: 10.1016/j.jesp.2009.11.004] [Medline: 20228874]
17. Hagiwara N, Slatcher RB, Eggly S, Penner LA. Physician racial bias and word use during racially discordant medical interactions. Health Commun 2017 Apr;32(4):401-408. [doi: 10.1080/10410236.2016.1138389] [Medline: 27309596]
18. Crossley SA, Kyle K, McNamara DS. Sentiment Analysis and Social Cognition Engine (SEANCE): an automatic tool for sentiment, social cognition, and social-order analysis. Behav Res Methods 2017 Jun;49(3):803-821. [doi: 10.3758/s13428-016-0743-z] [Medline: 27193159]
19. Crossley SA, Skalicky S, Dascalu M. Moving beyond classic readability formulas: new methods and new models. J Res Read 2019 Nov;42(3-4):541-561. [doi: 10.1111/1467-9817.12283]
20. Hu M, Liu B. Mining and summarizing customer reviews. In: KDD '04: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Association for Computing Machinery; 2004:168-177. [doi: 10.1145/1014052.1014073]
21. Liu B, Hu M, Cheng J. Opinion observer: analyzing and comparing opinions on the web. In: WWW '05: Proceedings of the 14th International Conference on World Wide Web: Association for Computing Machinery; 2005:342-351. [doi: 10.1145/1060745.1060797]
22. Mohammad SM, Turney PD. Emotions evoked by common words and phrases: using Mechanical Turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text: Association for Computational Linguistics; 2010:26-34 URL: https://aclanthology.org/W10-0204/ [accessed 2024-05-10]
23. Stone PJ, Dunphy DC, Smith MS. The General Inquirer: A Computer System for Content Analysis: MIT Press; 1966.
24. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media 2014 May 16;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
25. Lasswell HD, Namenwirth J. The Lasswell Value Dictionary: Yale University Press; 1969.
26. Namenwirth J, Weber R. Dynamics of Culture: Allen & Unwin; 1987.
27. Scherer KR. What are emotions? and how can they be measured? Social Science Information 2005 Dec;44(4):695-729. [doi: 10.1177/0539018405058216]
28. Ng SH. Language-based discrimination: blatant and subtle forms. J Lang Soc Psychol 2007 Jun;26(2):106-122. [doi: 10.1177/0261927X07300074]
29. Li Z, Chen MY, Banerjee J. Using corpus analyses to help address the DIF interpretation: gender differences in standardized writing assessment. Front Psychol 2020 Jun 3;11:1088. [doi: 10.3389/fpsyg.2020.01088] [Medline: 32581944]
30. Blair IV, Steiner JF, Fairclough DL, et al. Clinicians' implicit ethnic/racial bias and perceptions of care among Black and Latino patients. Ann Fam Med 2013;11(1):43-52. [doi: 10.1370/afm.1442] [Medline: 23319505]
31. Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. J Gen Intern Med 2013 Nov;28(11):1504-1510. [doi: 10.1007/s11606-013-2441-1] [Medline: 23576243]
32. Sabin JA, Greenwald AG. The influence of implicit bias on treatment recommendations for 4 common pediatric conditions: pain, urinary tract infection, attention deficit hyperactivity disorder, and asthma. Am J Public Health 2012 May;102(5):988-995. [doi: 10.2105/AJPH.2011.300621] [Medline: 22420817]
33. Sue DW, Capodilupo CM, Torino GC, et al. Racial microaggressions in everyday life: implications for clinical practice. Am Psychol 2007;62(4):271-286. [doi: 10.1037/0003-066X.62.4.271] [Medline: 17516773]

34. Statistics about diabetes. American Diabetes Association. URL: https://diabetes.org/about-us/statistics/about-diabetes [accessed 2023-06-28]
35. Ambulatory care use and physician office visits. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/fastats/physician-visits.htm [accessed 2023-06-28]
36. Fast facts on U.S. hospitals. American Hospital Association. URL: https://www.aha.org/statistics/fast-facts-us-hospitals [accessed 2023-06-28]
37. Cambria E, Havasi C, Hussain A. SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis. In: Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2012): AAAI Press; 2012:202-207 URL: https://cdn.aaai.org/ocs/4411/4411-21497-1-PB.pdf [accessed 2024-05-10]
38. Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. JAMA Netw Open 2022 Jan 4;5(1):e2144967. [doi: 10.1001/jamanetworkopen.2021.44967] [Medline: 35084481]
39. Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record. Health Aff (Millwood) 2022 Feb;41(2):203-211. [doi: 10.1377/hlthaff.2021.01423] [Medline: 35044842]
40. Barcelona V, Scharp D, Idnay BR, et al. A qualitative analysis of stigmatizing language in birth admission clinical notes. Nurs Inq 2023 Jul;30(3):e12557. [doi: 10.1111/nin.12557] [Medline: 37073504]
41. Goddu PA, O'Conor KJ, Lanzkron S, et al. Do words matter? stigmatizing language and the transmission of bias in the medical record. J Gen Intern Med 2018 May;33(5):685-691. [doi: 10.1007/s11606-017-4289-2] [Medline: 29374357]
42. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician use of stigmatizing language in patient medical records. JAMA Netw Open 2021 Jul 1;4(7):e2117052. [doi: 10.1001/jamanetworkopen.2021.17052] [Medline: 34259849]
43. Comprehensive diabetes care (CDC). National Committee for Quality Assurance. URL: http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2016-table-of-contents/diabetes-care [accessed 2017-11-29]
44. Ho YX, Gadd CS, Kohorst KL, Rosenbloom ST. A qualitative analysis evaluating the purposes and practices of clinical documentation. Appl Clin Inform 2014 Feb 26;5(1):153-168. [doi: 10.4338/ACI-2013-10-RA-0081] [Medline: 24734130]
45. Weis JM, Levy PC. Copy, paste, and cloned notes in electronic health records. Chest 2014 Mar;145(3):632-638. [doi: 10.1378/chest.13-0886] [Medline: 27845637]
46. Sukhera J, Wodzinski M, Rehman M, Gonzalez CM. The implicit association test in health professions education: a meta-narrative review. Perspect Med Educ 2019 Oct;8(5):267-275. [doi: 10.1007/s40037-019-00533-8] [Medline: 31535290]
47. van Ryn M, Hardeman R, Phelan SM, et al. Medical school experiences associated with change in implicit racial bias among 3547 students: a medical student CHANGES study report. J Gen Intern Med 2015 Dec;30(12):1748-1756. [doi: 10.1007/s11606-015-3447-7] [Medline: 26129779]
48. von Hippel W, Sekaquaptewa D, Vargas P. The linguistic intergroup bias as an implicit indicator of prejudice. J Exp Soc Psychol 1997 Sep;33(5):490-509. [doi: 10.1006/jesp.1997.1332]

Abbreviations

EHR: electronic health record
ICC: intraclass correlation
LIWC: Linguistic Inquiry and Word Count
PCA: principal component analysis
SEANCE: Sentiment Analysis and Social Cognition Engine

Edited by C Lovis; submitted 30.06.23; peer-reviewed by B Senst, M Chatzimina, X Jing; revised version received 26.09.23; accepted 23.04.24; published 23.05.24.

Please cite as:
Bilotta I, Tonidandel S, Liaw WR, King E, Carvajal DN, Taylor A, Thamby J, Xiang Y, Tao C, Hansen M
Examining Linguistic Differences in Electronic Health Records for Diverse Patients With Diabetes: Natural Language Processing Analysis
JMIR Med Inform 2024;12:e50428
URL: https://medinform.jmir.org/2024/1/e50428
doi:10.2196/50428

A Scalable Pseudonymization Tool for Rapid Deployment in Large Biomedical Research Networks: Development and Evaluation Study

Hammam Abu Attieh[1], MSc; Diogo Telmo Neves[1], BSc; Mariana Guedes[2,3,4], MSc, MD; Massimo Mirandola[5], PhD; Chiara Dellacasa[6], MSc; Elisa Rossi[6], MSc; Fabian Prasser[1], Prof Dr

1

2

3

4

5

6

Corresponding Author:
Hammam Abu Attieh, MSc

Abstract

Background: The SARS-CoV-2 pandemic has demonstrated once again that rapid collaborative research is essential for the future of biomedicine. Large research networks are needed to collect, share, and reuse data and biosamples to generate collaborative evidence. However, setting up such networks is often complex and time-consuming, as common tools and policies are needed to ensure interoperability and the required flows of data and samples, especially for handling personal data and the associated data protection issues. In biomedical research, pseudonymization detaches directly identifying details from biomedical data and biosamples and connects them using secure identifiers, the so-called pseudonyms. This protects privacy by design but allows the necessary linkage and reidentification.

Objective: Although pseudonymization is used in almost every biomedical study, there are currently no pseudonymization tools that can be rapidly deployed across many institutions. Moreover, using centralized services is often not possible, for example, when data are reused and consent for this type of data processing is lacking. We present the ORCHESTRA Pseudonymization Tool (OPT), developed under the umbrella of the ORCHESTRA consortium, which faced exactly these challenges when it came to rapidly establishing a large-scale research network in the context of the rapid pandemic response in Europe.

Methods: To overcome challenges caused by the heterogeneity of IT infrastructures across institutions, the OPT was developed based on programmable runtime environments available at practically every institution: office suites. The software is highly configurable and provides many features, from subject and biosample registration to record linkage and the printing of machine-readable codes for labeling biosample tubes. Special care has been taken to ensure that the algorithms implemented are efficient so that the OPT can be used to pseudonymize large data sets, which we demonstrate through a comprehensive evaluation.

Results: The OPT is available for Microsoft Office and LibreOffice, so it can be deployed on Windows, Linux, and MacOS. It provides multiuser support and is configurable to meet the needs of different types of research projects. Within the ORCHESTRA research network, the OPT has been successfully deployed at 13 institutions in 11 countries in Europe and beyond. As of June 2023, the software manages data about more than 30,000 subjects and 15,000 biosamples. Over 10,000 labels have been printed. The results of our experimental evaluation show that the OPT offers practical response times for all major functionalities, pseudonymizing 100,000 subjects in 10 seconds using Microsoft Excel and in 54 seconds using LibreOffice.

Conclusions: Innovative solutions are needed to make the process of establishing large research networks more efficient. The OPT, which leverages the runtime environment of common office suites, can be used to rapidly deploy pseudonymization and biosample management capabilities across research networks. The tool is highly configurable and available as open-source software.

(JMIR Med Inform 2024;12:e49646) doi:10.2196/49646

KEYWORDS

biomedical research; research network; data sharing; data protection; privacy; pseudonymization

Introduction

Background

As a response to the SARS-CoV-2 pandemic, many research projects have been rapidly set up to study the virus, its impact, and possible interventions [1,2]. This accelerated the general trend toward large collaborative networks in biomedical research [3,4]. These are motivated by the need to generate sufficiently large data sets and collections of biosamples, which are essential for developing new methods of personalized medicine and generating real-world evidence [5]. However, setting up such networks usually takes quite some time, as common tools and policies are needed to achieve interoperability and enable the required flows of data and biosamples [6,7]. One area in which this challenge is frequently encountered is the handling of personal data and the related data protection issues, which can arise in all processing steps, from collection [8] to sharing [9] and even analysis and visualization [10].

Laws and regulations, such as the European Union General Data Protection Regulation (GDPR) [11] or the US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [12], advocate for various strategies for the protection of personal data. In general terms, the GDPR prohibits the processing of sensitive categories of personal data, including medical data, unless consent is given. However, under certain conditions, processing is also possible without consent if technical and organizational safeguards are implemented [13]. Although there is no consensus on which protection methods are best suited for use in biomedical research [14], pseudonymization (also called coding or pseudo-anonymization) [15] is a common strategy, which can also be used to deidentify data under the HIPAA Privacy Rule. Pseudonymization is an essential aspect of the GDPR, as it is mentioned in multiple articles, in particular as a data minimization measure [16]. In this privacy-by-design approach, directly identifying data about study subjects are stored separately from biomedical data and biosamples, which are needed for scientific analyses [17]. The link between the different types of data and assets is established through secure identifiers, the so-called pseudonyms [18], which enable data linkage and allow the reidentification of subjects only if strictly necessary, for example, for follow-up data collection.

Objective

Although pseudonymization is done in almost any biomedical study, there are currently no pseudonymization tools that can rapidly be rolled out across many institutions. Existing tools, such as the Generic Pseudonym Administration Service (gPAS) [19] and Mainzelliste [20], are client-server applications, requiring server components to be deployed to and integrated into the institutions' IT infrastructures. Although this can have some important advantages (see the *Limitations and Future Work* section), it is usually time-consuming, for example, due to a lack of resources or efforts required to ensure compliance with local security policies. Moreover, using central services, such as the European Unified Patient Identity Management (EUPID) [21], is often not an option, for example, when data should be reused and consent is missing for this type of processing [22].

In this paper, we present the ORCHESTRA Pseudonymization Tool (OPT) that has been developed under the umbrella of the ORCHESTRA consortium. This project faced the challenges described in the previous paragraph when quickly establishing a large-scale research network as part of Europe's rapid pandemic response [23]. Hence, the OPT has been developed with the aim of supporting (1) the registration, pseudonymization, and management of study subject identities as well as biosamples; (2) rapid rollout across research network partners; and (3) scalability and simple configurability. The objective of this paper is to describe the design and implementation of the OPT and to offer insights into its usability and scalability, as evidenced by its deployment in the ORCHESTRA research network.

Methods

Ethical Considerations

The work described in this article covers the design and implementation of a generic research tool, which did not involve research on humans or human specimens and no epidemiological research with personal data. Therefore, no approval was required according to the statutes of the Ethics Committee of the Faculty of Medicine at Charité - Universitätsmedizin Berlin. However, the individual studies which use the tool usually have to apply for ethics approval. For example, the COVID HOME study within the ORCHESTRA project was approved by the Medical Ethical Review Committee of the University Medical Center Groningen (UMCG) under vote number METc 2020/158.

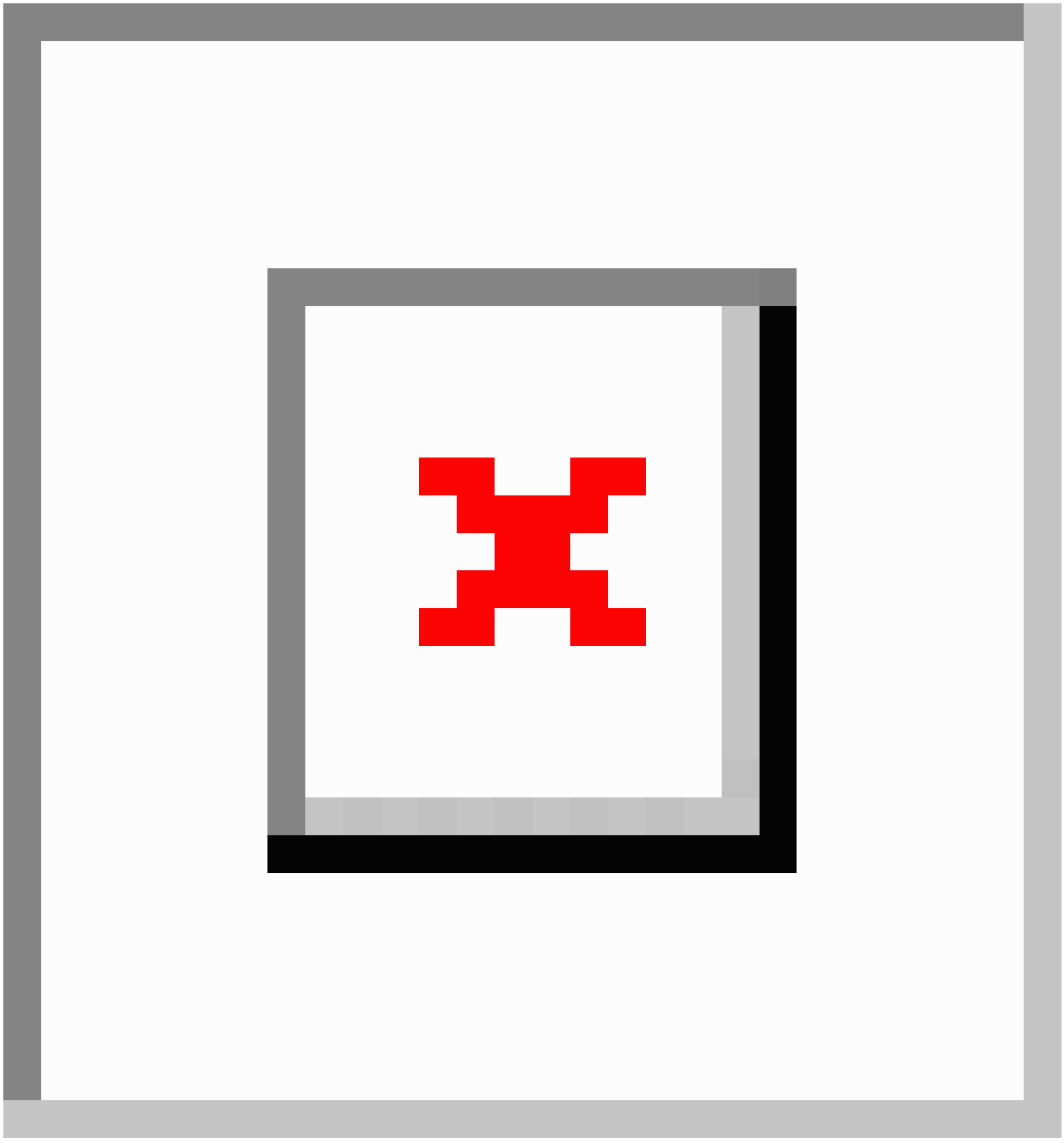
General Approach

The OPT has been designed to support general pseudonymization workflows that are needed in most biomedical research projects, as illustrated in Figure 1.

When a subject is admitted to the hospital, visits a study center, or has a follow-up visit, they are enrolled in the study. In this setting, the physicians or study nurses collect directly identifying and medical data and, according to the study protocol, the appropriate biosamples. The identifying attributes are entered into the OPT to create a unique pseudonym: the OPT Subject ID. During the follow-up visits, the study staff can use the OPT to retrieve an existing pseudonym from a subject that was already enrolled in the study. In all downstream data collection or processing, the OPT Subject ID can be used instead of identifying data so that the medical data are protected but still linked to the study subject and across visits. In addition, biosample data can also be entered into the OPT and linked to the appropriate subject to generate 1 or more additional pseudonyms: the OPT Biosample IDs. A label can then be generated for each biosample vial, containing the OPT Biosample ID, the OPT Subject ID, a DataMatrix Code, a QR code, or a barcode (containing the OPT Biosample ID) for tracking the biosample via scanners commonly used in laboratories. Study-specific information, for example, the exact information to capture for each study subject and biosample, the number and schedule of visits, and the types and schedules of biosample collections, can all be configured in the OPT. Moreover, in addition to its applicability in prospective studies, as described above, the software also supports importing existing

data about subjects and biosamples that can be used in retrospective study designs.

Figure 1. Basic concept of the OPT. IDAT: identifying data; MDAT: medical data; PID: patient ID; PSN: subject pseudonym; PSN-S: sample pseudonym; SDAT: sample data; SID: sample ID.



Implementation Details

To overcome challenges caused by the heterogeneity of IT infrastructures across different institutions and a potential lack of support by IT departments due to resource constraints, the OPT has been implemented based on programmable runtime environments that are available at practically any institution: office suites. These suites, especially the one by Microsoft, are among the most important and widely used applications around the world and still play a key role in many sectors today. The OPT is available for Microsoft Office as an Excel application and for LibreOffice as a Calc application. The application logic has been implemented in the embedded Basic scripting language using efficient algorithms for data management. Although Visual Basic for Applications is supported by Microsoft Office and LibreOffice Basic is supported by LibreOffice, they share similarities but are not fully compatible with each other. In the development process of the OPT, the Excel version serves as the primary implementation, and changes as well as additions are regularly ported to the LibreOffice version to achieve feature parity.

For generating the labels for the biosample vials, the OPT is delivered together with a single-page label printing application

that takes pseudonyms and metadata (eg, visit labels) as input and generates printable labels. Although this application is implemented using web technologies such as HTML, CSS, and JavaScript, it is delivered as files and can be executed locally without access to the internet. The label printing application works in any common web browser and can be called via the OPT. Properties of the labels to be printed can either be automatically transmitted via the URL for a single label or manually copied into the application via an input field for bulk printing of a larger number of labels. It is also possible to host the application on a web server. However, in this case, the URL function will be deactivated in the OPT to ensure that no data are sent to the server that hosts the application. It is important to note that the application still runs completely locally in the browser of the user, and no data ever leave the devices used to print labels. The pseudonyms and biosample metadata will be temporarily managed in the browser of the device.

Specific Functionalities

In addition to study subject and biosample management, the OPT also provides import and export functionalities, statistics, and a range of configuration options. In this section, we will briefly introduce each function, whereas a structured overview can be found in Multimedia Appendix 1. Regarding the subject-related functions, the OPT supports individual or bulk registration and a search function for finding pseudonyms for already registered subjects. An important feature of the software is a search function, required for any new patient or sample registration, which prevents multiple registrations of the same study participant. The search, to be performed as the first step of the registration, is linked to several data quality checks as well as a fuzzy record linkage process that prevents duplicate registrations. The bulk registration functionality enables the use of the OPT for retrospective pseudonymization of existing data sets. The search function supports wildcards and fuzzy matching across a configured set of master data attributes. Additional properties for the registered individuals can be documented to account for site-specific requirements.

Biosample-related functions are designed analogously to those for study subject management. In addition, labels can be generated and printed through the service described in the previous section.

Import and export functionalities are provided to enable the creation of backups (see the next section) and the migration from old versions of the OPT as part of update processes.

Finally, separate worksheets display statistical information about the data captured, such as the number of subjects registered or pseudonyms created for different study visits. Extensive configuration options are also available through a separate worksheet.

All functionalities of the OPT are described briefly in an integrated Quick User Guide and in detail in a comprehensive user manual [24].

Security Considerations and Features

The data collected during study subject and biosample registration, as well as the pseudonyms generated, are sensitive and a critical part of the data managed in any study. Hence, the confidentiality, integrity, and availability [25] of the data managed in the OPT must be ensured. In this context, the approach taken by the OPT clearly trades off some of the guarantees that could be provided by a client-server application against the possibility of rapid deployment and rollout. However, as described in the user manual, care has been taken to provide robust guarantees by specifying requirements on how the OPT should be deployed and used [24]. First, the OPT should not be placed on a local drive but on a network share that is integrated with the institution's Authentication and Authorization Infrastructure and, hence, provides means for controlling who is able to access the software in read or write mode and from which devices. Second, it is highly recommended that this share be backed up regularly so that data can be restored in case of problems. This should be complemented by regular, for example, daily, manual backups through the export functionality provided by the OPT and according to reminders that are displayed by the software. Finally, the office suites used as runtime environments do not provide multiuser support, and the application can only be opened by 1 user with write permission at any point in time. To enable parallel read access, the OPT comes with a script that opens a temporary read-only copy of the software. This allows, for example, laboratory technicians to use the OPT for generating biosample labels in parallel with ongoing registration processes. The measures described in this section have proven to be effective, and no problems have been encountered to date during extensive use of the software at many institutions (see the *Results* section).

Results

Overview of the Application

The graphical user interface of the OPT is divided into 10 different perspectives that provide access to the functionalities described in the previous sections. One of those sheets, the configuration sheet, is hidden from the users. All other sheets have write protection using the integrated protection functions of the spreadsheet software, except the input fields and the buttons, to ensure that data management is only performed through the specific functionalities provided by the software. A password is set by default for the write protection, which can be changed by the administrator at any time. However, it is important to keep the password safe. Figure 2 provides an overview of 4 important perspectives.

Figure 2. Perspectives of the OPT for (A) configuration, (B) registration and search, (C) data overview, and (D) statistics. OPT: ORCHESTRA Pseudonymization Tool.

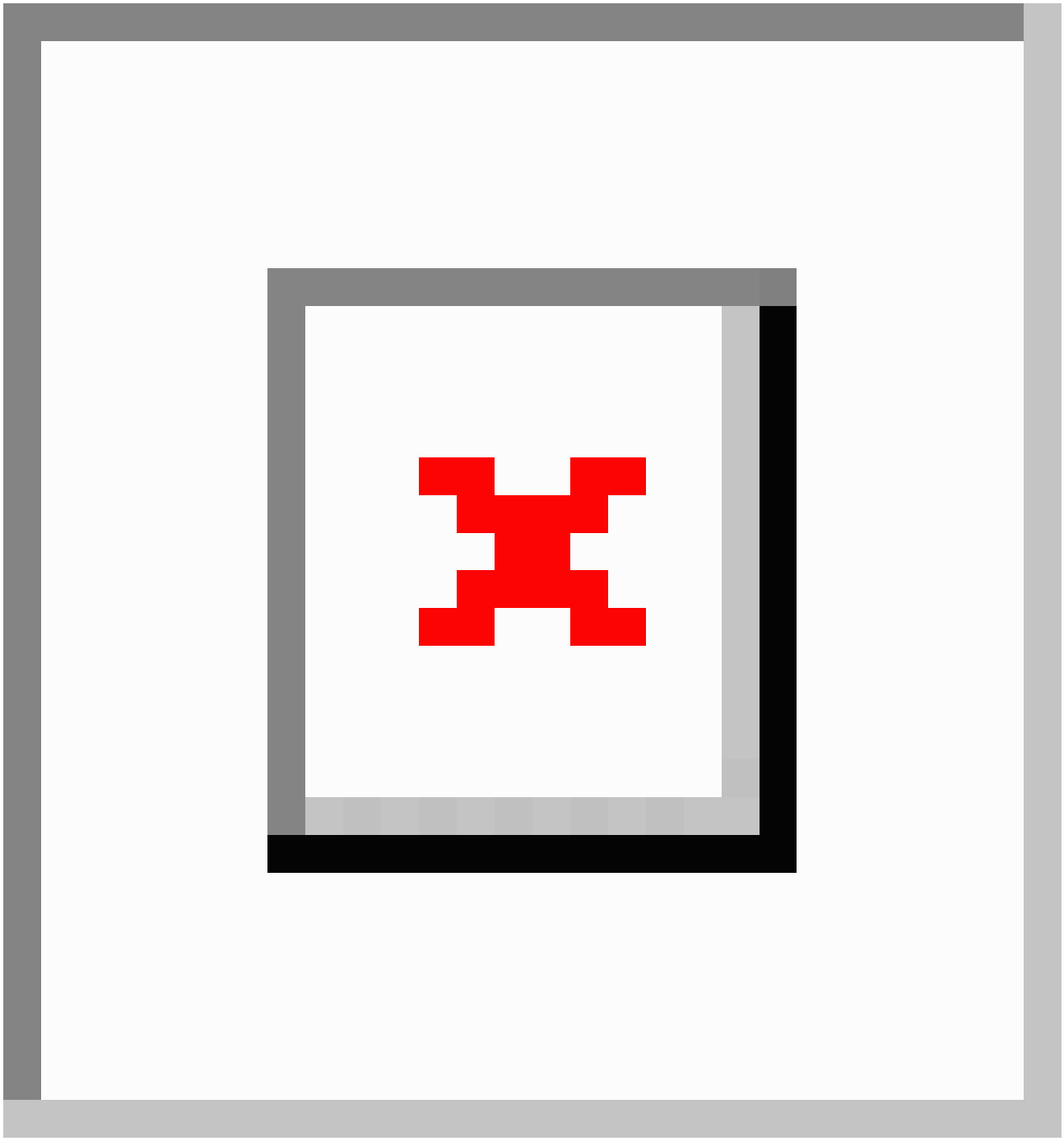
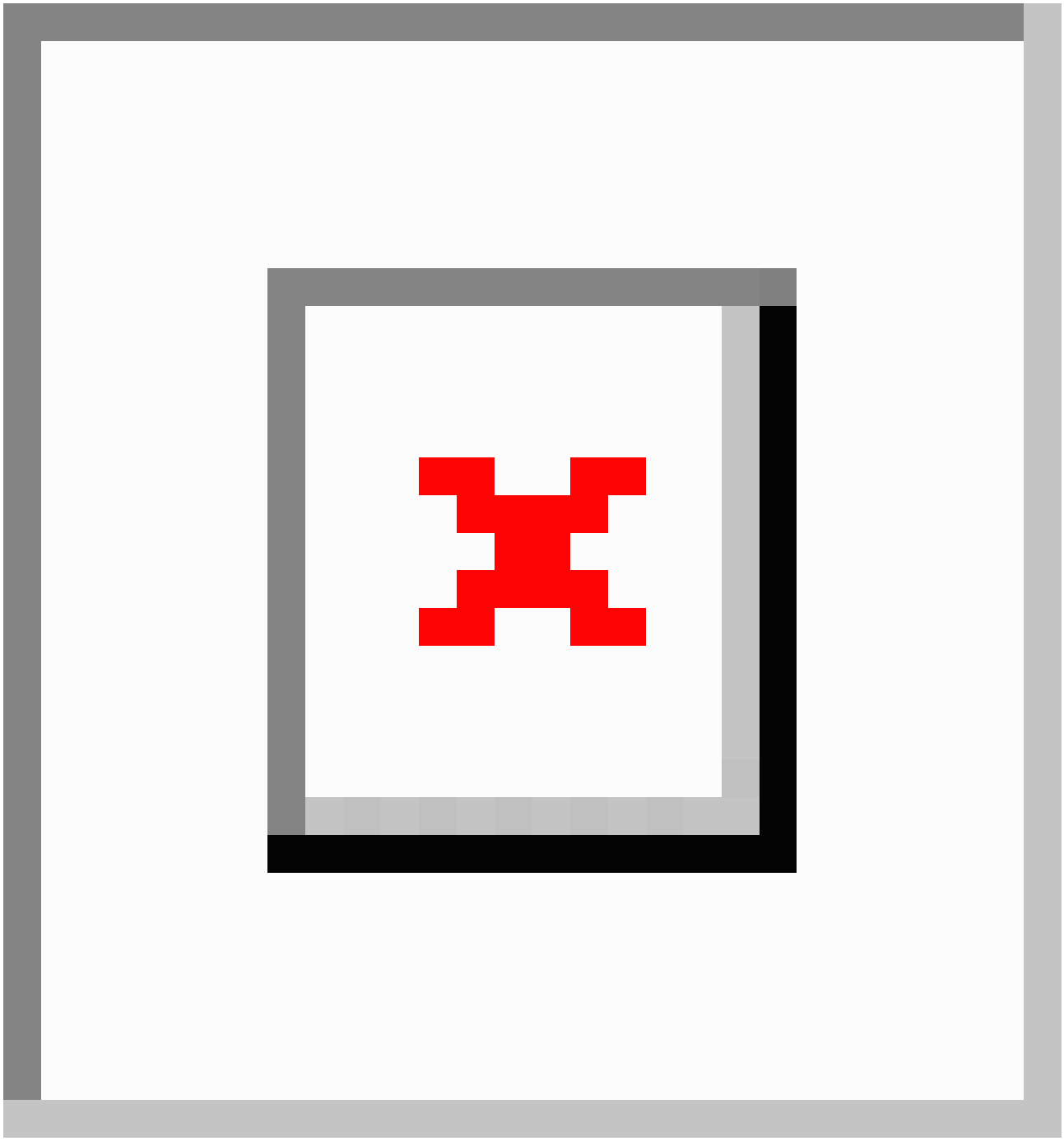


Figure 2A shows the configuration sheet, in which the specifics of the algorithm for generating pseudonyms, the study schedule, and the data fields to be documented can be specified. Figure 2B shows the interface provided for searching and registering subjects, with a search form on the left side of the sheet and a results list on the right side. All study subject data stored in the OPT are listed in the sheet shown in Figure 2C. This sheet also allows users to document any additional data that a site may require. Finally, Figure 2D shows a sheet providing statistical information on the number of subjects and biosamples registered, as well as insights into how these numbers have developed over time.

An overview of the label printing application is provided in Figure 3. As shown in the figure, the data that are to be printed on the labels are listed, and the number of rows and columns can be configured to support printing in bulk or for individual labels. The figure also shows an example of a sheet that can be printed and a detailed image of a single label. The data that are printed on those labels include the biosample and study subject IDs, the associated visit of the study schedule, and the biosample type.

Figure 3. Overview of the label printing application. OPT: ORCHESTRA Pseudonymization Tool.



Use of the OPT in the ORCHESTRA Project

ORCHESTRA is a 3-year international research project about the COVID-19 pandemic that was established in December 2020, involving 26 partners from 15 countries. The aim of ORCHESTRA is to share and analyze data from several retrospective and prospective studies to provide rigorous evidence for improving the prevention and treatment of COVID-19 and to better prepare for future pandemics [26,27].

The data management architecture in ORCHESTRA consists of 3 layers that build upon each other. The first layer is formed by "National Data Providers," which consist of the participating partners (universities, hospitals, and research networks). These provide the subject data and samples for joint analyses. On the second layer, "National Hubs" pool pseudonymized data in national instances of the Research Electronic Data Capture (REDCap) system [28]. Finally, the "ORCHESTRA Data Portal" forms the third layer, in which access to aggregated data and results is provided through a central repository.

In ORCHESTRA, the OPT was used for implementing pseudonymization at the data providers' sites. Each participating site named 1 or 2 persons responsible for technical aspects, such as setting up the required network share and installing updates, as well as several study nurses or clinicians, who would use the OPT. With these users, we performed regular training sessions and provided contact details in case of questions. As of June 2023, 19 instances of the OPT have been rolled out to 13 sites in 11 countries, including Germany, France, Italy, and Slovakia

in Europe; Congo in Africa; and Argentina in South America. A world map highlighting all the countries in which the OPT has been rolled out can be found in Multimedia Appendix 2.

On average, each instance of the OPT was used by up to 4 staff members. The OPT has been successfully rolled out, used, and maintained at large sites with committed IT departments, as well as at smaller, resource-constrained institutions. Overall, it has been in constant production use for more than 2 years. In the majority of the sites (10/13, 77%), the OPT Microsoft Excel version was used, whereas the remaining sites (3/13, 23%) used the LibreOffice release. In total, more than 10,000 study subjects and 15,000 samples have been registered in the OPT across all sites, and more than 10,000 labels have been printed. To evaluate the usability of the OPT, we conducted a survey among all active users, leveraging the widespread System Usability Scale [29] questionnaire, which includes 10 Likert-scale questions. During this survey, our system was designed to prevent multiple responses from individual participants and the submission of incomplete responses. We received 6 responses from 9 invited users, resulting in a score of 75 on a scale from 0 to 100, which adjectively translates to “good” [30].

Performance Evaluation

As mentioned, the OPT has been carefully designed to provide acceptable performance, even when large data sets are being processed or a large number of subjects or samples are being managed. In this section, we present the results of a brief performance evaluation. Our test environment consisted of an average office laptop, which was equipped with a quad-core 1.8 GHz Intel Core i7 CPU and a 64-bit Microsoft Windows 10 operating system. On top of it, Microsoft Excel 2016 (x32) and LibreOffice 7.0 (x64) were installed. Figure 4 provides an overview of the execution times of the most important functionalities of the OPT for different cohort sizes.

The numbers clearly show that the OPT works well and provides excellent performance for small or medium-sized data sets and acceptable performance for large data sets.

Figure 4. Execution times of the most important operations of the ORCHESTRA Pseudonymization Tool: (A) import, (B) registration, and (C) search.

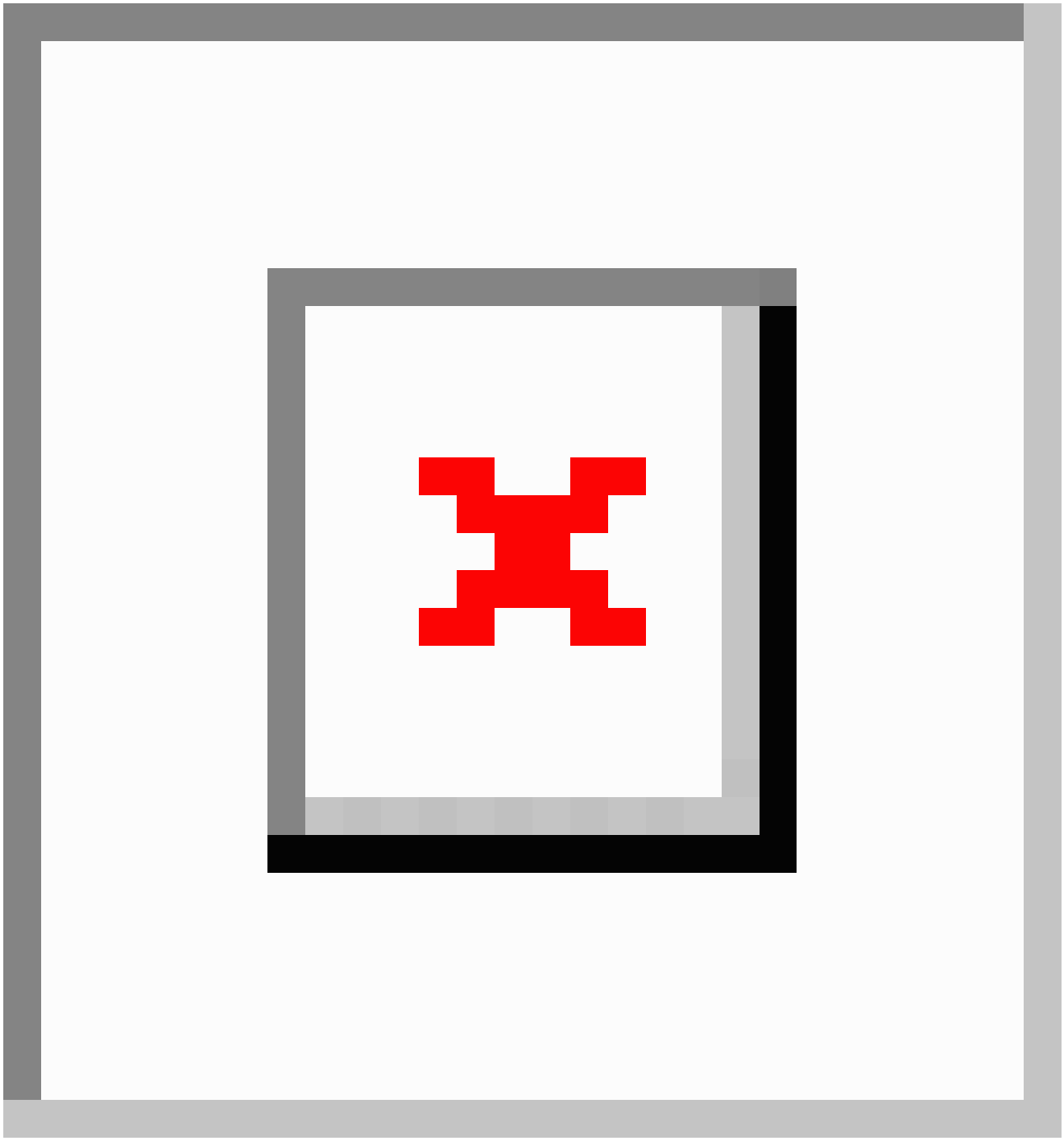


Figure 4A shows the average execution times for importing data about study subjects and samples. Data about subjects were imported into a completely empty OPT, whereas data about samples were imported into an OPT that already had the corresponding study subjects registered, so that each biosample was assigned to exactly 1 subject. For example, importing the data of 100,000 subjects took about 10 seconds in the Excel version and 54 seconds in the LibreOffice version. During the registration, the existence of the associated study subject in the OPT is checked, which makes the registration of samples slower compared to the registration of subjects. This is also noticeable in Figure 4B, which shows the average execution times for registering a single study subject or sample. As can be seen, using an OPT data set in which 100,000 entities were already registered, this took between 2 and 4 seconds in the Excel version and between 4 and 6 seconds in the LibreOffice version. Figure 4C shows the average execution times for searching for entities and obtaining their pseudonym, which is roughly twice as fast as the registration operation.

As performance is associated linearly with the number of entities already managed, subsecond response times can be expected for instances in which around 15,000 or fewer subjects or samples have been registered. This is consistent with our experiences from the deployments in the ORCHESTRA research network.

Discussion

Principal Findings

In this paper, we presented the OPT, a comprehensive, scalable, and pragmatic pseudonymization tool that can be rapidly rolled out across large research networks. To achieve this, the software has been implemented based on runtime environments that are available at practically any institution: office suites. The software supports a broad range of functionalities, from registering and pseudonymizing subject and biosample identities to search and depseudonymization functions, statistics about the data managed, as well as import and export features. We have described measures that are recommended to ensure the security of the data managed by the OPT and reported on our experiences gained after 2 years of successful operation in a large research network on COVID-19. Finally, we have also presented the results of a performance evaluation showing that the software provides excellent performance for small or medium-sized data sets and acceptable performance for large data sets. The OPT is available as open-source software [31] and can be configured to meet the needs of a wide range of biomedical research projects.

Limitations and Future Work

To achieve the design goals of the OPT, some compromises had to be made regarding data management. Compared to using client-server applications that use database management systems to store data, it is more difficult to ensure the confidentiality, integrity, and availability of the data managed with the OPT. There is also limited support for multiuser scenarios. However, we have developed and documented a set of measures that, if taken, help to still ensure a high level of data security. For this to work, it is important that users adhere to those recommendations. Therefore, all users of the OPT should familiarize themselves with the manual [24], and ideally, they should also be trained in the use and operation of the software. Despite these limitations, we strongly believe that our approach offers an innovative take on pseudonymization tools that can rapidly be rolled out across large research networks. Of course, it would be even more desirable if global standards for pseudonymization functions could be developed and agreed upon. Such global standards would ensure that solutions already existing at many research institutions are interoperable and can readily be used in joint research activities.

Comparison With Related Work

A range of pseudonymization tools has been described in the literature and are available as open-source software. However, they are either based on a client-server architecture and hence require quite some effort to be rolled out across sites, based on central services and hence not usable if consent is lacking for this type of processing, or offered as command-line utilities or programming libraries for IT experts.

Examples of client-server approaches include the work by Lablans et al [20] to provide a RESTful interface to pseudonymization services in modern web applications, which is based on a concept suggested by Pommerening et al [6] in 2006. Moreover, researchers from the University of Greifswald in Germany have designed and developed several client-server tools that can be used to manage subjects, samples, and other aspects of biomedical studies [32,33].

Examples of central services for pseudonymization include the EUPID, which was developed in 2014 by the Austrian Institute of Technology for the European Network for Cancer Research in Children and Adolescents project [21]. Another example is the Secure Privacy-preserving Identity management in Distributed Environments for Research (SPIDER) service, which was launched in May 2022 by the Joint Research Centre [34]. Both services support linking and transferring subject data across registries without revealing their identities. However, biosample data management is not possible with them. Further centralized concepts include the one described by Angelow et al [35].

Examples of command-line utilities, application programming interfaces, and programming libraries include the generic solution for record linkage of special categories of personal data developed by Fischer et al [36]; that by Preciado-Marquez et al [37]; and the PID (patient ID) generator developed by the TMF (Technologies, Methods and Infrastructure for Networked Medical Research e.V.), the German umbrella association for networked medical research [6].

Conclusion

Widely available office suites provide runtime environments that offer opportunities to rapidly roll out software components for biomedical studies across a wide range of large and resource-constrained research institutions. We have demonstrated this through the development, practical use, and evaluation of the OPT, which offers pseudonymization functionalities for study subjects and biosamples. As we believe that the software is of interest to the larger research community, it has been made available under a permissive open-source license [31].

Acknowledgments

This work has been funded by the European Union’s Horizon 2020 research and innovation programme under the project ORCHESTRA (grant agreement 101016167).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of the ORCHESTRA Pseudonymization Tool functions.
[PNG File, 233 KB - medinform_v12i1e49646_app1.png]

Multimedia Appendix 2
Map of countries in which the ORCHESTRA Pseudonymization Tool has been rolled out.
[PNG File, 229 KB - medinform_v12i1e49646_app2.png]

References

1. Dron L, Dillman A, Zoratti MJ, Haggstrom J, Mills EJ, Park JJH. Clinical trial data sharing for COVID-19-related research. J Med Internet Res 2021 Mar 12;23(3):e26718. [doi: 10.2196/26718] [Medline: 33684053]
2. R&D Blueprint. A coordinated global research roadmap: 2019 novel coronavirus. : World Health Organization; 2020 Mar 12 URL: https://www.who.int/publications/m/item/a-coordinated-global-research-roadmap [accessed 2024-04-12]
3. Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. Nat Biotechnol 2018 May 9;36(5):391-392. [doi: 10.1038/nbt.4128] [Medline: 29734317]
4. Walport M, Brest P. Sharing research data to improve public health. Lancet 2011 Feb 12;377(9765):537-539. [doi: 10.1016/S0140-6736(10)62234-9] [Medline: 21216456]
5. Mahmoud A, Ahlborn B, Mansmann U, Reinhardt I. Clientside pseudonymization with trusted third-party using modern web technology. Stud Health Technol Inform 2021 May 27;281:496-497. [doi: 10.3233/SHTI210212] [Medline: 34042618]
6. Pommerening K, Schröder M, Petrov D, Schlösser-Faßbender M, Semler SC, Drepper J. Pseudonymization service and data custodians in medical research networks and biobanks. In: INFORMATIK 2006 – INFORMATIK für Menschen: Gesellschaft für Informatik e.V; 2006, Vol. 1:715-721.
7. Tacconelli E, Gorska A, Carrara E, et al. Challenges of data sharing in European COVID-19 projects: a learning opportunity for advancing pandemic preparedness and response. Lancet Reg Health Eur 2022 Oct;21:100467. [doi: 10.1016/j.lanepe.2022.100467] [Medline: 35942201]
8. Rumbold J, Pierscionek B. Contextual anonymization for secondary use of big data in biomedical research: proposal for an anonymization matrix. JMIR Med Inform 2018 Nov 22;6(4):e47. [doi: 10.2196/medinform.7096] [Medline: 30467101]
9. Aamot H, Kohl CD, Richter D, Knaup-Gregori P. Pseudonymization of patient identifiers for translational research. BMC Med Inform Decis Mak 2013 Jul 24;13:75. [doi: 10.1186/1472-6947-13-75] [Medline: 23883409]
10. Wu X, Wang H, Zhang Y, Li R. A secure visual framework for multi-index protection evaluation in networks. Digit Commun Netw 2023 Apr;9(2):327-336. [doi: 10.1016/j.dcan.2022.05.007]
11. Regulation (EU) 2016/679 of the European Parliament and of the Council. Official Journal of the European Union. 2016 Apr 27. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679 [accessed 2024-04-12]
12. U.S. Department of Health and Human Services, Office for Civil Rights. HIPAA administrative simplification: regulation text: 45 CFR parts 160, 162, and 164 (unofficial version, as amended through March 26, 2013). U.S. Department of Health and Human Services. 2013 Mar 26. URL: https://www.hhs.gov/sites/default/files/hipaa-simplification-201303.pdf [accessed 2024-04-12]
13. Quinn P. Research under the GDPR - a level playing field for public and private sector research? Life Sci Soc Policy 2021 Mar 1;17(1):4. [doi: 10.1186/s40504-021-00111-z] [Medline: 33648586]
14. Rodriguez A, Tuck C, Dozier MF, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: a scoping review. Clin Trials 2022 Aug;19(4):452-463. [doi: 10.1177/17407745221087469] [Medline: 35730910]
15. Kohlmayer F, Lautenschläger R, Prasser F. Pseudonymization for research data collection: is the juice worth the squeeze? BMC Med Inform Decis Mak 2019 Sep 4;19(1):178. [doi: 10.1186/s12911-019-0905-x] [Medline: 31484555]
16. Gruschka N, Mavroeidis V, Vishi K, Jensen M. Privacy issues and data protection in big data: a case study analysis under GDPR. Presented at: 2018 IEEE International Conference on Big Data (Big Data); Dec 10 to 13, 2018; Seattle, WA p. 5027-5033. [doi: 10.1109/BigData.2018.8622621]
17. Lautenschläger R, Kohlmayer F, Prasser F, Kuhn KA. A generic solution for web-based management of pseudonymized data. BMC Med Inform Decis Mak 2015 Nov 30;15:100. [doi: 10.1186/s12911-015-0222-y] [Medline: 26621059]
18. European Union Agency for Cybersecurity, Drogkaris P, Bourka A. Recommendations on shaping technology according to GDPR provisions - an overview on data pseudonymisation. : European Network and Information Security Agency; 2018. [doi: 10.2824/74954]
19. Bialke M, Bahls T, Havemann C, et al. MOSAIC--a modular approach to data management in epidemiological studies. Methods Inf Med 2015;54(4):364-371. [doi: 10.3414/ME14-01-0133] [Medline: 26196494]
20. Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in modern web applications. BMC Med Inform Decis Mak 2015 Feb 7;15:2. [doi: 10.1186/s12911-014-0123-5] [Medline: 25656224]
21. Nitzlnader M, Schreier G. Patient identity management for secondary use of biomedical research data in a distributed computing environment. Stud Health Technol Inform 2014;198:211-218. [Medline: 24825705]

22. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015 Mar 20;350:h1139. [doi: 10.1136/bmj.h1139] [Medline: 25794882]
23. Connecting European cohorts to increase common and effective response to SARS-CoV-2 pandemic: ORCHESTRA. European Commission. 2022 Apr 21. URL: https://cordis.europa.eu/project/id/101016167/de [accessed 2023-06-02]
24. BIH-MI/opt: ORCHESTRA pseudonymization tool - user manual. GitHub. 2023 Sep 24. URL: https://github.com/BIH-MI/opt/blob/main/development/documentation/user-manual.pdf [accessed 2023-09-26]
25. ISO/IEC 27001:2022 information security, cybersecurity and privacy protection - information security management systems - requirements. : International Organization for Standardization; 2022 URL: https://www.iso.org/standard/27001 [accessed 2024-04-12]
26. Azzini AM, Canziani LM, Davis RJ, et al. How European research projects can support vaccination strategies: the case of the ORCHESTRA project for SARS-CoV-2. Vaccines (Basel) 2023 Aug 14;11(8):1361. [doi: 10.3390/vaccines11081361] [Medline: 37631929]
27. ORCHESTRA - EU horizon 2020 cohort to tackle COVID-19 internationally. ORCHESTRA. 2022 Sep 19. URL: https://orchestra-cohort.eu/ [accessed 2023-04-12]
28. 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]
29. Brooke J. SUS: a quick and dirty usability scale. In: Usability Evaluation in Industry: CRC Press; 1996:189-194.
30. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009 May;4(3):114-123 [FREE Full text]
31. BIH-MI/opt: ORCHESTRA pseudonymization tool. GitHub. 2023 Jun 2. URL: https://github.com/BIH-MI/opt [accessed 2023-06-02]
32. Bialke M. Werkzeuggestützte Verfahren für die Realisierung einer Treuhandstelle im Rahmen des zentralen Datenmanagements in der epidemiologischen Forschung [Dissertation].: Universitätsmedizin der Ernst-Moritz-Arndt-Universität Greifswald; 2016 URL: https://d-nb.info/1124566945/34 [accessed 2024-04-12]
33. Bialke M, Penndorf P, Wegner T, et al. A workflow-driven approach to integrate generic software modules in a trusted third party. J Transl Med 2015 Jun 4;13:176. [doi: 10.1186/s12967-015-0545-6] [Medline: 26040848]
34. SPIDER pseudonymisation tool. European Commission. 2023 May 4. URL: https://eu-rd-platform.jrc.ec.europa.eu/spider/ [accessed 2023-06-02]
35. Angelow A, Schmidt M, Weitmann K, et al. Methods and implementation of a central biosample and data management in a three-centre clinical study. Comput Methods Programs Biomed 2008 Jul;91(1):82-90. [doi: 10.1016/j.cmpb.2008.02.002] [Medline: 18406002]
36. Fischer H, Röhrig R, Thiemann VS. Simple Batch Record Linkage System (SimBa) – a generic tool for record linkage of special categories of personal data in small networked research projects with distributed data sources: lessons learned from the Inno_RD project. In: Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. 64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS): German Medical Science GMS Publishing House; 2019. [doi: 10.3205/19gmds118]
37. Preciado-Marquez D, Becker L, Storck M, Greulich L, Dugas M, Brix TJ. MainzelHandler: a library for a simple integration and usage of the Mainzelliste. Stud Health Technol Inform 2021 May 27;281:233-237. [doi: 10.3233/SHTI210155] [Medline: 34042740]

Abbreviations

EUPID: European Unified Patient Identity Management
GDPR: General Data Protection Regulation
gPAS: Generic Pseudonym Administration Service
HIPAA: Health Insurance Portability and Accountability Act
OPT: ORCHESTRA Pseudonymization Tool
PID: patient ID
REDCap: Research Electronic Data Capture
SPIDER: Secure Privacy-preserving Identity management in Distributed Environments for Research
SUS: System Usability Scale
TMF: Technologies, Methods and Infrastructure for Networked Medical Research e.V.

Edited by C Lovis; submitted 06.06.23; peer-reviewed by J Scheibner, X Wu; revised version received 03.10.23; accepted 07.03.24; published 23.04.24.

Please cite as:
Abu Attieh H, Neves DT, Guedes M, Mirandola M, Dellacasa C, Rossi E, Prasser F
A Scalable Pseudonymization Tool for Rapid Deployment in Large Biomedical Research Networks: Development and Evaluation Study
JMIR Med Inform 2024;12:e49646
URL: https://medinform.jmir.org/2024/1/e49646
doi:10.2196/49646

Original Paper

User Preferences and Needs for Health Data Collection Using Research Electronic Data Capture: Survey Study

Hiral Soni[1], PhD; Julia Ivanova[1], PhD; Hattie Wilczewski[1], BS; Triton Ong[1], PhD; J Nalubega Ross[1], PhD; Alexandra Bailey[1], MS; Mollie Cummins[1,2], PhD; Janelle Barrera[1,3], MPH; Brian Bunnell[1,3], PhD; Brandon Welch[1,4], PhD

[1]Doxy.me Research, Doxy.me Inc, Charleston, SC, United States

[2]College of Nursing, University of Utah, Salt Lake City, UT, United States

[3]Department of Psychiatry and Behavioral Neurosciences, University of South Florida, Tampa, FL, United States

[4]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Hiral Soni, PhD
Doxy.me Research
Doxy.me Inc
18 Broad Street, 3rd Floor
Suite 6 and 7
Charleston, SC, 29401
United States
Phone: 1 8444369963
Email: sonihiralc@gmail.com

Abstract

Background: Self-administered web-based questionnaires are widely used to collect health data from patients and clinical research participants. REDCap (Research Electronic Data Capture; Vanderbilt University) is a global, secure web application for building and managing electronic data capture. Unfortunately, stakeholder needs and preferences of electronic data collection via REDCap have rarely been studied.

Objective: This study aims to survey REDCap researchers and administrators to assess their experience with REDCap, especially their perspectives on the advantages, challenges, and suggestions for the enhancement of REDCap as a data collection tool.

Methods: We conducted a web-based survey with representatives of REDCap member organizations in the United States. The survey captured information on respondent demographics, quality of patient-reported data collected via REDCap, patient experience of data collection with REDCap, and open-ended questions focusing on the advantages, challenges, and suggestions to enhance REDCap's data collection experience. Descriptive and inferential analysis measures were used to analyze quantitative data. Thematic analysis was used to analyze open-ended responses focusing on the advantages, disadvantages, and enhancements in data collection experience.

Results: A total of 207 respondents completed the survey. Respondents strongly agreed or agreed that the data collected via REDCap are accurate (188/207, 90.8%), reliable (182/207, 87.9%), and complete (166/207, 80.2%). More than half of respondents strongly agreed or agreed that patients find REDCap easy to use (165/207, 79.7%), could successfully complete tasks without help (151/207, 72.9%), and could do so in a timely manner (163/207, 78.7%). Thematic analysis of open-ended responses yielded 8 major themes: survey development, user experience, survey distribution, survey results, training and support, technology, security, and platform features. The user experience category included more than half of the advantage codes (307/594, 51.7% of codes); meanwhile, respondents reported higher challenges in survey development (169/516, 32.8% of codes), also suggesting the highest enhancement suggestions for the category (162/439, 36.9% of codes).

Conclusions: Respondents indicated that REDCap is a valued, low-cost, secure resource for clinical research data collection. REDCap's data collection experience was generally positive among clinical research and care staff members and patients. However, with the advancements in data collection technologies and the availability of modern, intuitive, and mobile-friendly data collection interfaces, there is a critical opportunity to enhance the REDCap experience to meet the needs of researchers and patients.

(JMIR Med Inform 2024;12:e49785) doi:10.2196/49785

KEYWORDS

Research Electronic Data Capture; REDCap; user experience; electronic data collection; health data; personal health information; clinical research; mobile phone

Introduction

Background

Accurate and complete health outcome data directly from patients or study participants (hereon referred to as *patients*) are critical for health care and research [1-3]. Unfortunately, it can be burdensome to extract patient-reported health data that researchers or providers need [4,5]. Collecting patient-reported outcomes data is becoming increasingly important in clinical research and care [6,7]. Self-administered web-based questionnaires, which patients can complete at a clinic or at home, are becoming a conventional approach to collect data for clinical research. Web-based questionnaires have advantages of being low-cost and easy to deploy at scale. A variety of clinical research electronic data capture (EDC) tools exist to streamline remote data collection and management. These systems comply with privacy regulations, integrate with different tools (such as electronic health records [EHRs]) for efficient data collection, and reduce the effort of sharing data [8]. However, user experience, cost, and maintenance of such commercial EDC systems are often prohibitive. An understanding of user experiences and preferences regarding EDC tools is critical in assessing stakeholder needs, satisfaction, and challenges in clinical and research settings.

REDCap (Research Electronic Data Capture; Vanderbilt University) is a global, secure web application for building and managing EDC for clinical research [9,10]. Developed by Vanderbilt University, REDCap is freely available for its consortium members (ie, network of nonprofit collaborators and supporters), who have an established agreement with the university. REDCap is compliant with global privacy regulations (such as the Health Insurance Portability and Accountability Act [HIPAA] of 1996) and used by more than 2.2 million researchers in more than 140 countries [9]. REDCap allows researchers to build and conduct electronic surveys, track and manage study information, schedule visits, and manage databases that are fully customizable and at no cost [11]. REDCap is designed to support data capture for research studies, providing (1) an intuitive interface for validated data capture, (2) audit trails for tracking data manipulation and export procedures, (3) automated export procedures for seamless data downloads to common statistical packages, and (4) procedures for data integration and interoperability with external sources.

Although REDCap is widely used, user needs and preferences of EDC via REDCap have rarely been studied [12,13]. For example, 1 usability study of a REDCap-based patient-facing intervention reported that patient participants found REDCap useful and easy to use but showed concerns about wordiness and inconsistent visual design [13]. Researchers have reported frequently on the implementation, use, and interventions using REDCap [10,14-20]. Understanding the preferences and needs of REDCap administrators and researchers using REDCap to capture data could help enhance existing features and EDC processes in general. While REDCap is a robust clinical research data management system, this study solely focuses on the experience of REDCap as an EDC tool. To the best of our knowledge, such preferences have not yet been studied.

Objective

The aim of this study was to survey REDCap administrators and researchers in the United States to assess their experience with REDCap, including perspectives on advantages, challenges, and suggestions for enhancement.

Methods

Study Settings and Respondents

We conducted a web-based survey with representatives of member organizations listed as REDCap Partners on the REDCap website [21]. The roles of the listed members were unclear at the time of invitation sent via email. The email communication included information related to the study goals, voluntary participation, and a link to the REDCap survey. Respondents were compensated with a US $10 electronic gift card for completing the survey.

Ethical Considerations

This study was reviewed and approved as exempt human subjects research by the Medical University of South Carolina Institutional Review Board (Pro00082875).

Survey Design

We developed a web-based survey with multiple-choice and free-response questions (Multimedia Appendix 1) to capture the perspectives of researchers and administrators from participating REDCap consortium organizations. Our research team includes experts in biomedical informatics, behavioral sciences, mixed methods research, and user experience. The survey included 4 sections, as follows:

- *Demographics*: multiple-choice questions capturing participant role in their respective organization (Q1) and organization use of REDCap (Q2)
- *Quality of patient-reported data collected via REDCap*: Likert-scale questions capturing perspectives (ranging from 1=strongly agree to 5=strongly disagree) on the accuracy, reliability and completeness of data reported using REDCap (Q3)
- *Patient experience with REDCap*: Likert-scale question focusing on perspectives (ranging from 1=strongly agree to 5=strongly disagree) on REDCap usability, including ease of use, success rate, and completion time (Q4).
- *Data collection experience*: Free-response questions asking about the advantages (Q5), challenges (Q6), and suggestions of enhancements related to data collection, patient experience, and engagement (Q7).

Data Collection and Analysis

We collected and managed study data using REDCap EDC tools hosted at the Medical University of South Carolina [22,23]. We generated plots and univariate statistics to summarize the data (eg, frequencies, means, SDs, and percentages). We conducted 1-way ANOVA tests to determine differences in data quality and patient experience variables by participant role and REDCap use duration. For the ANOVAs, the primary role variable was restructured to include "Educators" in the "Other" category due to the low sample size (n=1). Excel (Microsoft Corp) and SPSS (version 29; IBM Corp) were used for analyses. Free-response questions were qualitatively analyzed to identify emerging themes related to REDCap data collection experience [24]. We randomly selected 15% of the responses for initial coding and codebook development. The coding unit was done by the entirety of the participant entry. Thematic analysis of all qualitative data was done over 4 iterations using MAXQDA, during which emergent themes were identified. While the research team reviewed and honed the codes and codebook, 1 team member coded and finalized thematic coding. Discrepancies were resolved through consensus. Emergent themes were organized by frequency and topic, allowing for further qualitative analysis using complex coding query to determine concurrent themes. We reported the total frequencies per code, which may not align with the number of participants. For example, 1 participant may report a code multiple times throughout their response [25]. While thematic analysis allows us to identify principle emergent themes, it also can help identify uncommon trends that may be significant but would require further investigation in follow-up research [26]. Responses from incomplete surveys with missing quantitative or qualitative responses were excluded from the analysis

Results

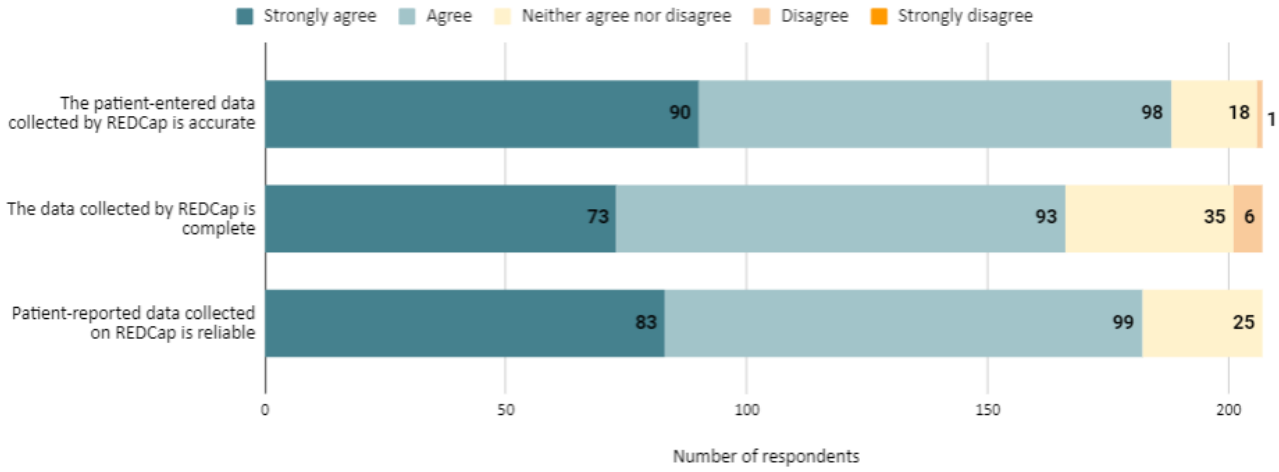
Demographics

Between October and November 2020, 3058 representatives from 1676 REDCap member organizations in the United States were invited to complete the survey. In total, 285 (9.3%) invitees started the survey, of which 207 completed the survey. Most (150/207, 72.5%) respondents were REDCap administrators, followed by researchers (25/207, 12.1%). Furthermore, 1 (0.5%) respondent was an educator and 31 (15%) respondents served in other roles, including IT directors and managers, research coordinators and managers, program managers, project managers, director of research, library directors, and data analysts. Respondents reported that their organization had used REDCap for <5 years (92/207, 44.4%), 5 to 10 years (83/207, 40.1%), or >10 years (32/207, 15.5%).

Quality of Patient-Reported Data Collected via REDCap

We asked respondents about their perspectives of the quality of the survey data, including the accuracy, reliability, and completeness of the data collected using REDCap (Figure 1). Most respondents strongly agreed or agreed that the data collected via REDCap are accurate (188/207, 90.8%), reliable (182/207, 87.9%), and complete (166/207, 80.2%). We observed no statistically significant group differences in accuracy ($F_{2,204}$=1.003; P=.37), completeness ($F_{2,204}$=0.243; P=.78), or reliability ($F_{2,204}$=0.245; P=.78) among respondent role groups. Furthermore, we observed no statistically significant group differences in accuracy ($F_{2,204}$=0.672; P=.51), completeness ($F_{2,204}$=0.045; P=.96), or reliability ($F_{2,204}$=1.712; P=.18) among REDCap use groups.

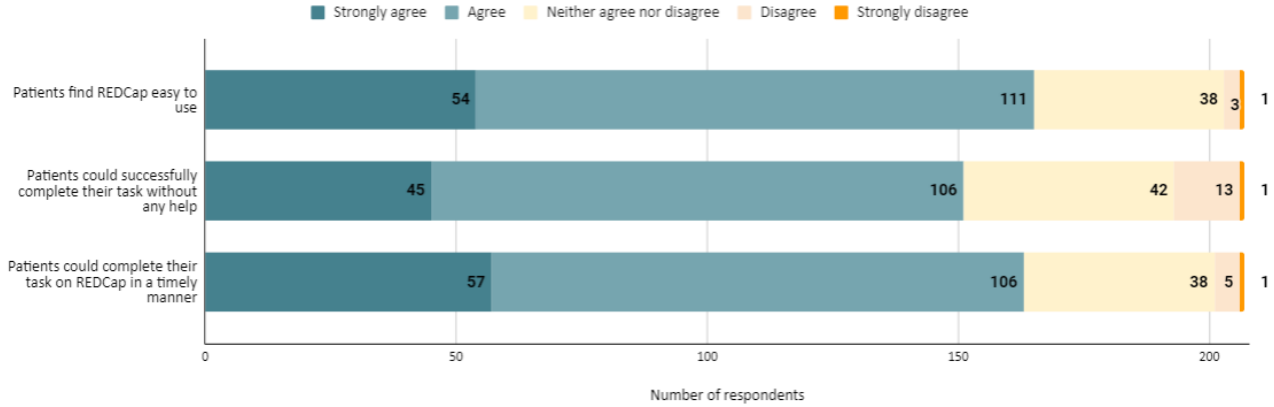
Figure 1. Quality of patient-reported data collected via REDCap (Research Electronic Data Capture).



Patient Experience With REDCap

We also asked respondents about their perspectives on patient experiences with completing surveys using REDCap. Figure 2 summarizes their responses. More than half of respondents strongly agreed or agreed that patients find REDCap easy to use (165/207, 79.7%), could successfully complete tasks without help (151/207, 72.9%), and could do so in a timely manner (163/207, 78.7%). We observed no statistically significant group differences in ease ($F_{2,204}$=2.025; P=.13), successful task completion ($F_{2,204}$=0.671; P=.51), or timely task completion ($F_{2,204}$=2.303; P=.10) among respondent role groups. Furthermore, we observed no statistically significant group differences in ease ($F_{2,204}$=0.711; P=.49), successful task completion ($F_{2,204}$=1.851; P=.16), or timely task completion ($F_{2,204}$=2.000; P=.13) among REDCap user groups.

Figure 2. Patient experience with REDCap (Research Electronic Data Capture).



REDCap Advantages, Challenges, and Enhancement Suggestions

We asked respondents about the advantages, challenges, and suggestions for future enhancements using free-response questions. The analysis yielded 8 primary codes: survey development, user experience, survey distribution, survey results, training and support, technology, security, and platform features. Within each of these themes, responses were further categorized at secondary and tertiary levels. Multimedia Appendix 2 shows the qualitative codebook with illustrative examples for each code. Table 1 shows the frequencies response classification of advantages, disadvantages, and enhancements for each code category based on respondents' responses.

Table 1. Counts and percentages of response classification of REDCap (Research Electronic Data Capture) users[a].

Code	Advantages, n (%)	Challenges, n (%)	Enhancements, n (%)
Survey development			
Design	52 (50.5)	59 (47.6)	40 (47.1)
Survey design	10 (9.7)	39 (31.5)	21 (24.7)
Response and logic	13 (12.6)	22 (17.7)	18 (21.2)
Survey setup	20 (19.4)	2 (1.6)	0 (0)
Flexibility	7 (6.8)	1 (0.8)	3 (3.5)
Organization	1 (1)	1 (0.8)	0 (0)
Testing	0 (0)	0 (0)	3 (3.5)
Customizations	7 (100)	17 (65.4)	15 (65.2)
Language support	0 (0)	9 (34.6)	8 (34.8)
Project interactions	0 (0)	3 (100)	6 (100)
Feature suggestions	1 (100)	16 (100)	49 (100)
User experience			
Usability	105 (55.9)	29 (61.7)	8 (53.3)
Ease of use	57 (30.3)	7 (14.9)	1 (6.7)
Accessibility	1 (0.5)	1 (2.1)	3 (20)
Intuitiveness	3 (1.6)	2 (4.3)	1 (6.7)
User-friendliness	11 (5.9)	4 (8.5)	2 (13.3)
Reliability	3 (1.6)	0 (0)	0 (0)
Simplicity	8 (4.3)	4 (8.5)	0 (0)
User interface	9 (52.9)	27 (50.9)	33 (52.4)
Visual interface	2 (11.8)	5 (9.4)	29 (46)
Devices	5 (29.4)	9 (17)	0 (0)
Functionality	1 (5.9)	8 (15.1)	0 (0)
Design configuration	0 (0)	4 (7.5)	1 (1.6)
Mobile experience	20 (60.6)	16 (64)	21 (50)
Ease of use	3 (9.1)	0 (0)	0 (0)
Interface	4 (12.1)	5 (20)	3 (7.1)
Mobile friendly	5 (15.2)	0 (0)	3 (7.1)
Mobile app	1 (3)	4 (16)	15 (35.7)
Patient experien ce	34 (55.7)	13 (48.1)	5 (62.5)
Convenience	10 (16.4)	0 (0)	0 (0)
Engagement	9 (14.8)	6 (22.2)	3 (37.5)
Patient input	3 (4.9)	8 (29.6)	0 (0)
Patient log-in	3 (4.9)	0 (0)	0 (0)
Efficiency	1 (1.6)	0 (0)	0 (0)
Empowerment	1 (1.6)	0 (0)	0 (0)
Researcher experience	8 (100)	0 (0)	0 (0)
Survey distribution and reminders			
Invitations and scheduling	20 (52.6)	25 (53.2)	25 (54.3)
Automated scheduling and messaging	5 (13.2)	1 (2.1)	2 (4.3)
Save and return	3 (7.9)	15 (31.9)	9 (19.6)

Code	Advantages, n (%)	Challenges, n (%)	Enhancements, n (%)
Invitation approaches	9 (23.7)	6 (12.8)	5 (10.9)
Calendar integration	1 (2.6)	0 (0)	3 (6.5)
Patient opt out	0 (0)	0 (0)	2 (4.3)
Reminders	7 (87.5)	7 (87.5)	7 (63.6)
Email text	0 (0)	0 (0)	1 (9.1)
Follow-up with patients	1 (12.5)	1 (12.5)	3 (27.3)
Easy distribution	11 (100)	0 (0)	3 (100)
Results and data			
Results view	5 (100)	0 (0)	4 (100)
Data sharing	8 (100)	0 (0)	3 (100)
Data quality	3 (100)	1 (100)	0 (0)
Training and support			
Education and training	7 (100)	7 (100)	8 (100)
Support	8 (100)	15 (100)	18 (100)
Patient support	2 (100)	12 (100)	16 (61.5)
Patient education and communication	0 (0)	0 (0)	10 (38.5)
Patient feedback	1 (100)	0 (0)	6 (100)
User misunderstanding and error	3 (100)	8 (100)	1 (100)
Technology and accessibility			
Consent	5 (100)	0 (0)	3 (100)
Technology integration	11 (100)	1 (100)	12 (100)
Technology access	17 (100)	51 (100)	0 (0)
Technology literacy	1 (100)	33 (100)	0 (0)
Security			
Privacy and compliance	15 (100)	2 (40)	2 (100)
Trust in technology	0 (0)	3 (60)	0 (0)
Platform features			
Data collection	14 (56)	4 (50)	5 (50)
Comprehensive	5 (20)	0 (0)	0 (0)
Data administration	3 (12)	1 (12.5)	3 (30)
Offline access	1 (4)	3 (37.5)	2 (20)
Familiarity	2 (8)	0 (0)	0 (0)
Cost	10 (100)	0 (0)	0 (0)
Comparison with other platforms	3 (100)	5 (100)	0 (0)
No input	7 (100)	22 (100)	52 (100)

[a]Due to the coding process (eg, double coding), the total number of secondary and tertiary codes may not add up to the primary code or 100%. The percentages are calculated based on the total number of codes in secondary and tertiary categories.

Survey Development and Customization

Respondents perceived that REDCap surveys were generally easy (20 codes) and quick (2 codes) to set up, build, organize, and maintain (2 codes). One participant commented on these topics, “Easy to build surveys Easy to make questions easy to answer Easy to build branching questions.”

However, respondents also noted that incorrect setup by the study staff and limited default formatting options and flexibility could be challenging in developing and completing surveys (3 codes).

While some respondents pointed out that REDCap provides continuous releases with new features (2 codes) and various design and automation options to ask a variety of questions for

efficient data collection (8 codes), respondents frequently pointed out the value of well-designed survey instruments in gathering high-quality information and engaging patients. They reported that complex, poorly designed surveys and ambiguous instructions (39 codes) could result in poor patient experience, potentially impacting the survey response rate and quality of data gathered. Respondents provided suggestions for enhancing survey design capabilities to streamline survey design and layout for the patients (including simplifying survey formatting, survey nesting abilities, and use of embedded fields). Respondents also suggested pilot testing of surveys before sending them out to patients (3 codes) and for study teams to follow best practices and guidelines to be more informed in survey methodologies and development. For example, 1 respondent commented:

> *Study teams following best practices with survey methodology and design, which can involve keeping surveys short & sweet, choosing appropriate field types for the question at hand, phrasing questions and response options well to reduce mental burden and make it easier for patients to answer questions.*

Respondents also reported that the availability of various response types, data validation, and branching logic ensure high-quality data collection (13 codes). One respondent commented on this advantage, "The wide array of validations can help patients enter data correctly."

Another respondent noted similarly, "Data validation and branching logic make participants conform to data standards and allows researchers to obtain higher quality data."

While data validation was discussed positively, respondents more frequently noted the challenges with response and logic types (22 codes), often pointing out that the actual response and logic types available from REDCap are not conducive to good survey design. One respondent made a clear reference to this issue saying, "It all depends on who sets up the survey, but until recently it has been a challenge to create grids of disparate data entry fields."

In addition, some respondents noted that due to the logic types, patients can make critical mistakes affecting the completeness of the data:

> *...branching logic at a very question to determine if they qualify or not. Sometimes, they accidently select different value in a hurry, and the survey gets completed. It is hard for them to change the response or refill the survey without admin help.*

Respondents noted many enhancement potentials within this category, such as voice input (4 codes), superior data entry experience (5 codes), use of a more conversational approach in response types (1 code), more effective multimedia (5 codes), and gamification of survey (2 codes). While REDCap offers multimedia options, respondents often suggested that options become more interactive and effective:

> *...more visual aids in questions, and the ability to answer with images. For example, by painting the areas afflicted on an image.*

One respondent explained how multimedia may be further useful:

> *...ability to add images to response options. Especially when working with minorities (traffic lights, or smiley faces).*

In addition to the design of the surveys, respondents noted that while REDCap surveys are readily customizable (7 codes), there are far more reported challenges (17 codes) and need for enhancements (15 codes). Respondents noted customization was not possible in some cases: "Default formatting options are limited."

However, many respondents focused on the lack of multi-language support (9 codes) as the critical challenge:

> *...multi-linguistic support. This is always a challenge for any software system/platform, and REDCap is no different...*

They frequently suggested enhancements to include multi-language support (8 codes) and customizations in forms' appearance (6 codes). For example, 1 respondent mentioned, "Allow for some more customization of the overall look/feel of surveys."

With respect to challenges with survey interactions, respondents reported that REDCap capabilities at the time did not send new surveys or allow patients to complete future surveys if previous surveys were incomplete (2 codes). One respondent mentioned the following:

> *...[t]he longitudinal design functionality in REDCap requires a participant to take each form before moving to the next, but our experiment design does not require this, and sometimes people will miss sessions and need to move on to the form for the next one. But if we stack all of the forms in one event, we cannot direct people to an individual form, only to the queue.*

One participant commented on REDCap's "inability to provide staff log-in status." (1 code). Respondents requested features for internal messaging or chat between study staff (2 codes), enhancing flow and cross-linking between projects (2 codes), ability to easily add study staff members outside of the organization (1 code), and ability for patients to skip longitudinal surveys (1 code).

User Experience

Respondents perceived REDCap to be easy to use for both patients (ie, to take surveys) and the study staff (ie, to build and distribute surveys; 57 codes). One respondent commented as follows:

> *REDCap is the easiest way to survey patients, families, and staff who are not part of our study team. We would not be able to conduct these surveys without it!*

They also perceived REDCap to be user-friendly (11 codes), simple (8 codes), intuitive (3 codes), timely (2 codes), and reliable (3 codes). Although some respondents reported REDCap allows for quick data collection (7 codes), they perceived that

lengthy or poorly designed surveys (eg, too many clicks and not enough instructions) could lead to fatigue and poor participation (15 codes). While the usability perceptions were generally positive, respondents reported that the platform was not as user-friendly or outdated as other commercial data collection platforms (4 codes), unintuitive (2 codes), and clunky for study staff (4 codes). They reported that "REDCap is not the simplest tool to learn how to use" for study staff (4 codes) and patients (3 codes). Respondents suggested the need to enhance accessibility features, such as the ability to change font size, screen reader view, and text-to-voice, among others (3 codes). In total, 8 (3.9%) of 207 respondents reported that the REDCap interface was advantageous for study staff considering its consistent interface and automated features, which reduce burden.

Respondents generally reported REDCap's visual user interface as challenging to use. Although some respondents perceived the interface to be clean or simple looking (9 codes) and optimized for various devices (5 codes), other respondents perceived that REDCap's interface was not modern looking (7 codes) or appealing (5 codes). One respondent mentioned, "The web interface of our survey pages are very basic, and narrow," whereas another respondent said, "[REDCap has] Very set layout of each item, can't make it look more 'modern' like other websites are at this time."

Respondents considered REDCap as not having a configurable design (4 codes) and some noted the user interface's poor functionality (8 codes). One respondent described both issues when explaining the challenges of the user interface:

> *REDCap is simply not user friendly in any way. The data structures are often too rigid and frankly outdated in being an effective tool for data collection.*

Respondents suggested the redesign of the REDCap user interface to be consistent with modern data collection platforms (27 codes), options to change the visual appearance and formatting of the surveys (3 codes), adding progress tracking aids (such as an automatic progress bar) for patients (2 codes), and a more flexible interface (1 code).

Some respondents appreciated REDCap's mobile access (4 codes), availability of mobile apps for study staff (REDCap mobile app; 2 codes) and patients (MyCap; 6 codes) supporting offline data collection, and perceived REDCap to be easy to use on mobile devices (3 codes) and mobile friendly (5 codes). While respondents appreciated the mobile interface, they reported that the mobile experience is affected by poor and suboptimal mobile user interface and scaling on smaller screens (5 codes). One participant reported the following:

> *We design our surveys on a computer, but many of our participants use their phones. We try to check how answers scale when the screen size changes, but some phones rescale to a different aspect ratio leading to challenges.*

They also reported that although the REDCap mobile app is available for study staff, it is not ideal and is difficult for study staff to set up the app (4 codes). One respondent mentioned the following:

> *I think that the REDCap mobile app is a bit too far separated from the web version, in as much as there is no access to external modules and other important features.*

Respondents suggested a need for an enhanced mobile app and interface (21 codes), including advanced capabilities for the study staff to view study records and perform analysis (2 codes) and push notifications (2 codes). One respondent mentioned the following:

> *[They need] better workflows with mobile phones, like notifications instead of just text messages. Something like an App except not the current one which is focus on asymmetric internet access.*

Respondents also commented on patient experience with REDCap. Overall, respondents noted that REDCap makes it easier for patients to complete the surveys at their convenience (10 codes), all while increasing engagement levels (9 codes). They saw REDCap as a way to make data collection more efficient and empowered (2 codes), especially as patients did not need to register or remember usernames or passwords to use the platform (3 codes). One participant said, "[Survey] Can be done at the patient's convenience from any digital device." A common challenge reported was the patient's desire and motivation to complete the surveys, being able to use the platform, and fatigue with lengthy surveys (13 codes). Suggestions for improving patient experience included maintaining engagement using visual aids and gamification (3 codes), a patient dashboard to keep them up to date on status of longitudinal studies (1 code) and making the platform more patient friendly (1 code). One respondent commented as follows:

> *For longer surveys, having a way of maintaining engagement by making the surveys more interactive (e.g. fun feedback to participants as they progress) would be nice. Some periodic messages of encouragement like "Great job!" "Keep it up!"*

Survey Distribution and Reminders

Respondents found it advantageous that REDCap included multiple ways to invite patients, such as emails or embedded links (4 codes). REDCap surveys were easy to distribute (11 codes) and could be automated and scheduled on a timeline easily. One participant commented on this aspect, "It can send surveys to participants directly, and on a schedule when the project is longitudinal." REDCap's ability to send patients custom links was an advantage respondents liked (3 codes): "For online surveys: able [to] send individualized email links...automated email with message that has piping upon completion." One respondent pointed out that there was "no scheduling component for visits" and suggested this feature. One respondent suggested the ability to send attachments with automatic notifications.

In addition, the ability to send completion reminder emails to patients was reported to reduce the burden on clinic staff while engaging patients (7 codes). Reminders also allowed the study staff members to follow up on incomplete surveys but 1 respondent mentioned that this was challenging while respondents suggested for improvements in customizing

reminders and enhanced tracking for incomplete surveys longitudinally (3 codes):

> *If there was a more efficient way to upload and manage patient invitations, as well as identify which patients have completed the survey within previous xx months therefore a new survey invitation does not need to be sent.*

Respondents noted patients sometimes missed invitations and reminders because email service providers blocked the emails (2 codes): "We have had email providers block REDCap emails, specifically Yahoo.com email." There was also confusion about the email sender as the emails were "from" REDCap instead of the study staff (2 codes):

> *From my experience... The emails that are sent out to respondents are not user friendly. The 'From' text box comes from REDCap, not from my email address.*

In addition, this respondent noted the emails were not user-friendly, sometimes arriving with broken links going to patient's junk mail, and requiring patients create a completely new log-in to complete a survey. One respondent suggested REDCap may "make it easier to send mass emails that are individually linked with the patient's profile; create a prettier or more visually appealing interface for patients." Furthermore, integrations to link communications to personal calendars were thought to be beneficial (3 codes). Respondents wanted a way to automatically opt out patients from surveys that were being distributed over a period (2 codes). One participant stated they, "would really like to be able to set a flag for opt-out subject [s] when distributing surveys over a period of time. We currently have to remove their emails to prevent future distribution."

Respondents commented on the "Save and Return" feature (25 codes), which allows patients to leave and return using a unique code to complete the survey at a later time. Although REDCap's Save and Return feature existed, respondents noted that this feature was often difficult to use (15 codes). They reported that patients may forget or not save their return code or may not know how to return to the survey, resulting in incomplete data or delay in data collection. One respondent commented, "It is not always obvious how to 'save and return later' if that is an option or even be aware that that is an option." Respondents suggested improvements (9 codes) to send the unique save and return code via emails, with reminders and save in invitation logs such that the study staff could provide it to patients if needed. In addition, respondents suggested that improving user-friendliness and patient awareness of this feature could increase response rates and data completion. A participant noted the following:

> *If they [patients] don't complete the survey the first time they often forget their return code and lose it. It would really help if the reminder emails had the return code, of if it could be included on the survery [sic] invitation log page that would make it much easier to find and give to the patient.*

Results and Data

Respondents liked that REDCap made data exportation easy for storage and analysis purposes (8 codes). Not only was it easy to export data out of the REDCap survey tool, it also made the analysis of the data much easier for the study staff, even those with minimal statistics training. As 1 respondent put it, "[REDCap has a] good translation into a dataset [and] easy statistics for those with minimal statistical training." Respondents (4 codes) pointed out the need for improving data exports and seamless communication with third-party solutions to send and receive information:

> *Being able to send to communicate and receive information from other software programs like Clinical Conductor for Demographic information and seamless data uploads.*

Respondents perceived that it is easy to create reports and monitor patient responses on REDCap and review specific data points (5 codes). Some respondents provided suggestions to edit charts and graphics as well as being able to share user- or survey-specific data (3 codes). For example, 1 respondent mentioned the following:

> *Ability for researchers to edit/modify graphics that can be automatically displayed with reports within redcap. This would facilitate researchers' ability to use those charts."*

Another respondent mentioned, "built in tools to share summary-level data (you vs the whole study) or findings." While respondents perceived that REDCap allows capturing accurate and complete high-quality data (3 codes), 1 respondent mentioned the following:

> *As with every self-service data entry portal accuracy of self-service data entry is wildly unreliable. There is real value to having a trained rep assisting the client enter information, when possible.*

Training and Support

Respondents reported that REDCap's active online community and support allowed REDCap users (including administrators and researchers) to find information and answers on how to manage, design, and conduct surveys (8 codes): "...it has a huge user base and a great consortium full of all the information you need to begin administering [surveys]." Respondents mentioned needing REDCap or IT support for patients to complete consent forms or surveys (12 codes). Although support existed for survey designers and administrators, it did not extend to patients completing surveys. Respondents suggested REDCap needed a way to educate or support patients in completing surveys (10 codes) and obtain help via on-demand messaging to study staff members (2 codes). As 1 participant suggested, REDCap should allow patients to "Click icon and get video explaining any information on a field." Another participant asked that REDCap have the following:

> *Dedicated instrument defined support button at the top that takes participants to a page made by the study team where we can put in a zoom room link monitored by study staff, phone numbers, or some pointers on definitions/examples on the instrument.*

Respondents also suggested a need for obtaining standardized patient feedback surveys to better engage them and understand their experience (6 codes).

Although some respondents mentioned REDCap required minimal training to get started (7 codes), some respondents (especially REDCap administrators) mentioned the need for training survey designers to set up REDCap tools and surveys to design high-quality surveys (7 codes). When asked about challenges, 1 participant mentioned the following:

> *Lack of resources for support (in person- phone) and functionality. It is not always easy and takes a lot of time to build tools. Not able to use to its fullest capacity or correctly—basically training ourselves. Library or community network does not help either. Not knowing how to set up properly more complicating functions inhibits usage.*

Respondents suggested more information and mandatory training for survey builders, including better guidelines and training videos to enhance builder and patient experience (8 codes). Respondents also perceived that patients taking surveys often do not understand how to fill out surveys or certain questions (8 codes) and having expert survey designers and well-designed surveys could alleviate these concerns (1 code).

Technology

Respondents often noted challenges of access to the internet and devices (51 codes) as well as technology literacy (33 codes):

> *Patients [without] a computer, device, or smart phone may not be able to use REDCap.*

As REDCap is web based, data collection could be difficult in rural and low-resource areas due to lack of access to technology (4 codes), such as a computer or reliable internet connection. Another participant noted, "I do work in global health, so our colleagues in resource-limited settings have challenges with the internet connection."

They also noted REDCap's ability to integrate with other technologies, such as messaging tools (eg, Twilio) as well as open application programmable interface to be beneficial (11 codes). In comparison, 1 participant noted as a challenge that, "integrating the ReCap [sic] extract with Epic [EHR] data. But once the system is setup it's easy to maintain." Respondents suggested integrations with other clinical trial management systems for seamless data transfers and EHRs to conduct surveys or autopopulate patient medical information:

> *The only other thing that would be super cool is if it could blow surveys into EPIC for documentation when needed.*

Respondents also referred to the informed consent capabilities of REDCap (8 codes). Even though they noted the consent module to be advantageous to obtain remote informed consents especially after the COVID-19 pandemic (5 codes), respondents suggested more enhancements, such as a 1-step consent process (3 codes).

Security

Respondents commented positively on the security and compliance of REDCap (15 codes). They reported that HIPAA compliance and the ability to store patient data securely are important advantages of REDCap. One participant commented that "all client data can be stored in one HIPAA compliant platform."

Respondents mentioned mistrust of technology (3 codes) could make patients feel uncomfortable sharing medical information on web-based platforms. One respondent commented that surveys requiring password protection are difficult for patients. They also provided enhancement suggestions (2 codes) related to maintaining HIPAA compliance, enhancing security, and assuring patients that their health information is safe and secure with REDCap.

Platform Features

Respondents found REDCap advantageous in enabling researchers to collect and patients to provide health data remotely (23 codes):"It has made it much easier for patients to submit their questionnaires and information using an online platform," especially during and after the COVID-19 pandemic.

Respondents perceived REDCap as a comprehensive or versatile (5 codes) data collection solution noting the following: "It provides us a comprehensive tool for collecting, tracking, and managing patient data and outreach." They also noted administration and maintenance (3 codes) to be advantageous as REDCap allows "being able to maintain administrative research tasks together with the data collection." They noted REDCap's offline data collection (using REDCap mobile apps) to be challenging (3 codes) and suggested that the offline feature should be improved for better data collection experience (2 codes). In addition, respondents noted the familiarity with REDCap among researchers (2 codes) and seamlessness (1 code) for the study personnel to be advantageous.

In addition, REDCap being available for free to REDCap consortium members was sought to be beneficial (10 codes). While some respondents noted REDCap being simpler and easier than other commercial platforms and paper forms (3 codes), some also noted that REDCap's interface was not easy to use or user-friendly compared to modern data collection tools (5 codes).

No Input

Respondents did not provide inputs with respect to advantages, disadvantages, and enhancement suggestions stating lack of experience or ability to provide inputs or not using REDCap for patient data collection (81 codes). Some nonsensical or unrelated comments lacking information context or irrelevant responses were excluded from the analysis. For example, when asked about enhancement suggestions for REDCap, 1 participant responded, "To REDCap or??."

Discussion

Overview

This study aimed to identify the advantages, challenges, and future opportunities for enhancements from the perspectives of REDCap administrators and researchers. To the best of our knowledge, this is one of the early studies of user perspectives on REDCap services and features. We believe that the findings of this study will aid REDCap developers and consortium users in better understanding stakeholder needs to enhance and customize REDCap features as well as researchers in improved survey development and data collection.

Principal Findings

Respondents had overwhelmingly positive perceptions of REDCap's survey design and data collection interface. The vast majority of respondents agreed or strongly agreed that data collected via REDCap were accurate (188/207, 90.8%), reliable (182/207, 87.9%), and complete (166/207, 80.2%). They found REDCap advantageous as it is free for its consortium members, secure, and easy to use. Respondents also perceived REDCap as easy and flexible to create and customize surveys including a variety of response and validation options, which make data collection easier for survey takers. However, respondents pointed out that poor survey design—often attributed to human factors (eg, lengthy forms and lack of knowledge among study staff) or technology limitations (eg, restrictions in survey and visual formatting in REDCap)—could result in poor patient experience and, ultimately, response and completion rates. Optimal design of survey forms is critical for assuring patient comprehension of the forms and accurate data collection [27,28]. Furthermore, direct investigations of REDCap user experiences and preferences could allow better understanding of the need for study staff and patient education. In addition, further research related to user needs for survey development and optimization can lead to enhancing their experience of developing high-quality surveys. One respondent pointed out the following:

> *It [REDCap] needs a much better understanding of how users engage the questions on a form (e.g., sit and watch users and staff try to figure out acceptable data type entries!). Needs a solid revamping in how it works "out front" and to run a series of user groups—patient and staff.*

Although respondents appreciated the availability of REDCap's community support for administrators and study staff, they pointed out that REDCap has room for improvement in this realm: the tool is not simple to learn, and there is a need for more training of study staff to help develop efficient, unambiguous survey instruments that can enhance patient experience. Poorly designed surveys and questions could potentially lead to incomplete responses and inaccurate data. Respondents pointed out the need for supporting patients, especially to ensure they understand the questions and can obtain help when needed in filling out surveys. Direct help from study staff members to fill out surveys or having the ability to directly send a message to study staff could alleviate misunderstandings and errors in completing surveys. Previous research suggests that the ability to obtain clarifications about survey questions can enhance response accuracy [28]. Further research and availability of resources are necessary to guide study staff members in creating well-designed instruments. In addition, understanding the factors affecting patients' experience in completing REDCap surveys and reasons for misunderstanding and errors could also enhance the REDCap experience and health data collection processes.

Opinions on patient experience and usability were more mixed. Most respondents agreed or strongly agreed that patients found REDCap easy to use (90.4%), able to be completed without assistance (79.8%), and able to be completed in a timely manner (87.5%). These strongly positive perceptions of REDCap usability are consistent with a prior study in which 6 out of 7 participants needed no help using REDCap, achieved 71% to 100% task completion, and provided 89% positive reaction words [13]. Qualitative outcomes showed that respondents perceived REDCap made it convenient for patients to provide data remotely without having to log in or remember credentials. Although they commented that patients can complete REDCap surveys using a device of choice (such as a laptop or mobile), technology access and technology literacy appeared to be a concern. Living in rural or low-income areas also presented issues for survey access. Respondents noted low-resource areas without stable internet access meant data collection was not reliable. Lack of internet access not only meant surveys could not be accessed but also meant the data collection process could be interrupted. REDCap's MyCap and REDCap Mobile app can allow study staff and patients to complete the collection of data offline, but they were also deemed challenging due to the lack of features compared with the web interface. In a study by Doyle et al [19], the REDCap mobile interface was less favorably received by participants. Similarly, REDCap's *Save and Return* feature allows users to complete surveys at a later time, which could be helpful during poor internet access; however, respondents recommended enhancements in the feature to improve patient experience, specifically an easier way for patients to remember and retrieve the return code. One participant noted this difficulty that patients face in attempting to use the feature:

> *If they don't complete the survey the first time, they often forget their return code and lose it. It would really help if the reminder emails had the return code, or if it could be included on the survey invitation [log-in] page...*

It is imperative to better understand patient and research participant experience with REDCap in completing surveys via larger and direct studies.

This study identified opportunities to improve the usability of REDCap. Respondents suggested enhancements in the patient-facing survey user interface to be in line with present EDC tools on the market, wanting a sleeker, modern, and cleaner looking interface. A variety of EDC tools are available for health care and non–health care data collection providing modern, device-friendly, and intuitive user interfaces to promote patient engagement [29-32]. In recent years, virtual conversational agents or chatbots have emerged as intuitive and engaging mediums for data collection. Modern data collection tools allow

survey designers to develop chatbot-based interactions to collect health data mimicking human-to-human conversations. Studies have shown that individuals prefer chatbot-based conversational data collection experience in comparison to traditional web-based forms [33,34]. Visual and graphical enhancements in REDCap appearance of surveys, patient communication, and researcher interface could support modernization of REDCap-based surveys, thus providing study staff and patients with clear and effective experience of health data collection.

Respondents wanted the mobile interface updated to look more like other commercial products, such as Qualtrics or SurveyMonkey. As more individuals are using mobile devices to obtain health information, it is of great importance to enhance their experience with mobile data collection [35]. They also suggested that the mobile apps have similar features as the web-based REDCap. Other requests included REDCap to support more languages or a translation service, where surveys could be translated to patients' preferred languages. Though it has some language capabilities, including Spanish, respondents wanted more language options built into REDCap. In addition, there was concern about the literacy of patients leading to suggestions for REDCap to include tools allowing patients with various literacy levels to access surveys. Respondents suggested inclusion of voice capabilities and more multimedia and gamification features in response options, such as a picture interface where patients could locate their pain visually for researchers. Inclusion of these features could further enhance the experience among patients with higher accessibility needs and low literacy. We also noted that some respondents suggested features that were available within REDCap at the time of conducting the survey. Suggestions included availability of REDCap's mobile version, embedded fields for responses, and integrations with messaging services such as Twilio. This again points out the need for education among study staff and organizational administrators to enable the optimal and effective use of REDCap features.

Limitations

This study is not without limitations. Although we recruited over 200 respondents, the sample size is small in comparison with the existing user base. We recruited fewer researchers (25/207, 12.1%) than administrators (150/207, 72.5%) who may be more directly involved in survey design and data collection. We also did not ask for participants' training and experience with REDCap. Future studies should focus on better understanding user perspectives (especially researchers) while also considering the type and amount of REDCap training received by the user. We asked individuals' opinions that are valuable but may be subject to bias, incomplete recall, or lack of information. For example, we asked information about their institution's REDCap use, but we did not include a response option or decline responding if they did not have accurate information. We also used REDCap as the platform to conduct our survey, which may have potentially biased responses by familiarizing participants with REDCap more than necessary. Participants' free-ended responses may have been influenced by how our study was designed or how the features were used. Future, more direct studies are warranted to better understand preferences. We recruited respondents from current REDCap consortium members, who may be more likely to believe REDCap is highly usable, as they may act as REDCap champions within institutions. We may be missing critical information by not capturing the perspectives of people who are not frequent users or consortium members. Future research should capture opinions of novice or past REDCap users. We also did not ask for information about participants' institutions, REDCap versions and plug-ins used, or institutional policies and customizations. It is possible that participant feedback may be related to institutional requirements or policies. Furthermore, we asked researcher and administrator opinions on patient experience. However, we did not directly assess the patient or research participant experience. Understanding patient experience is important to study in future research. In addition, a comparison of the REDCap experience with other EDC platforms could provide a better understanding of study staff and patient needs. A recent study compared individuals' experience in completing health forms using REDCap versus a chatbot platform. The results revealed that over 69% of participants preferred a chatbot for data collection with higher usability and net promotor scores for the chatbot [33]. The chatbot provided superior engagement and interactivity and was perceived as more intuitive than a standard, web-based REDCap interface. Future studies should look into better understanding study staff and patient needs to optimize survey development and data collection experience.

Conclusions

This pilot study aimed to assess stakeholder perspectives on experience with REDCap as an electronic health data collection tool. The findings revealed researchers and administrators perceive REDCap as a valued, low-cost resource that enables them to remotely collect and report health data in a secure and easy way. They also indicated a generally positive health data collection experience by clinical research and care staff members and patients. Although, with the advancements in data collection technologies and availability of interactive and intuitive user interfaces, there is a critical opportunity to enhance the REDCap experience to meet the needs of its vast user base of researchers and patients.

Acknowledgments

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health (award number 1R41LM013419-01). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. BB was funded by the National Institute of Mental Health (grant K23MH118482). BW was funded by the National Library of Medicine (grant R41LM013419).

Conflicts of Interest

BW is a shareholder of Doxy.me Inc and Dokbot LLC. All other authors are employees of Doxy.me Inc, a commercial telemedicine company. All authors declare no other conflicts of interest.

Multimedia Appendix 1
Survey.
[DOCX File , 16 KB - medinform_v12i1e49785_app1.docx]

Multimedia Appendix 2
Qualitative codebook.
[DOCX File , 26 KB - medinform_v12i1e49785_app2.docx]

References

1. Caron-Flinterman JF, Broerse JE, Bunders JF. The experiential knowledge of patients: a new resource for biomedical research? Soc Sci Med 2005 Jun;60(11):2575-2584. [doi: 10.1016/j.socscimed.2004.11.023] [Medline: 15814182]
2. Hanley B, Truesdale A, King A, Elbourne D, Chalmers I. Involving consumers in designing, conducting, and interpreting randomised controlled trials: questionnaire survey. BMJ 2001 Mar 03;322(7285):519-523 [FREE Full text] [doi: 10.1136/bmj.322.7285.519] [Medline: 11230065]
3. Saczynski JS, McManus DD, Goldberg RJ. Commonly used data-collection approaches in clinical research. Am J Med 2013 Nov;126(11):946-950 [FREE Full text] [doi: 10.1016/j.amjmed.2013.04.016] [Medline: 24050485]
4. Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. Ann Transl Med 2018 Feb;6(3):42 [FREE Full text] [doi: 10.21037/atm.2018.01.13] [Medline: 29610734]
5. van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for research: an international comparative study. BMC Med Inform Decis Mak 2016 Jul 13;16(1):90 [FREE Full text] [doi: 10.1186/s12911-016-0332-1] [Medline: 27411943]
6. Sacristan JA, Aguaron A, Avendaño C, Garrido P, Carrion J, Gutierrez A, et al. Patient involvement in clinical research: why, when, and how. Patient Prefer Adherence 2016 Apr;10:631-640. [doi: 10.2147/ppa.s104259]
7. van der Scheer L, Garcia E, van der Laan AL, van der Burg S, Boenink M. The benefits of patient involvement for translational research. Health Care Anal 2017 Sep 24;25(3):225-241 [FREE Full text] [doi: 10.1007/s10728-014-0289-0] [Medline: 25537464]
8. Summary of the HIPAA privacy rule. U.S. Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html [accessed 2021-12-09]
9. REDCap homepage. REDCap. URL: https://www.project-redcap.org/ [accessed 2022-05-09]
10. Patridge EF, Bardyn TP. Research electronic data capture (REDCap). J Med Libr Assoc 2018 Jan 12;106(1) [FREE Full text] [doi: 10.5195/jmla.2018.319]
11. REDCap (Research Electronic Data Capture) - Harvard Catalyst. Harvard Catalyst. URL: https://catalyst.harvard.edu/redcap/ [accessed 2024-06-05]
12. Reichold M, Heß M, Kolominsky-Rabas P, Gräßel E, Prokosch HU. Usability evaluation of an offline electronic data capture app in a prospective multicenter dementia registry (digiDEM Bayern): mixed method study. JMIR Form Res 2021 Nov 03;5(11):e31649 [FREE Full text] [doi: 10.2196/31649] [Medline: 34730543]
13. Stambler DM, Feddema E, Riggins O, Campeau K, Breuch LA, Kessler MM, et al. REDCap delivery of a web-based intervention for patients with voice disorders: usability study. JMIR Hum Factors 2022 Mar 25;9(1):e26461 [FREE Full text] [doi: 10.2196/26461] [Medline: 35333191]
14. Kianersi S, Luetke M, Ludema C, Valenzuela A, Rosenberg M. Use of research electronic data capture (REDCap) in a COVID-19 randomized controlled trial: a practical example. BMC Med Res Methodol 2021 Aug 21;21(1):175 [FREE Full text] [doi: 10.1186/s12874-021-01362-2] [Medline: 34418958]
15. Tamuhla T, Tiffin N, Allie T. An e-consent framework for tiered informed consent for human genomic research in the global south, implemented as a REDCap template. BMC Med Ethics 2022 Nov 24;23(1):119 [FREE Full text] [doi: 10.1186/s12910-022-00860-2] [Medline: 36434585]
16. Wong TC, Captur G, Valeti U, Moon J, Schelbert EB. Feasibility of the REDCap platform for single center and collaborative multicenter CMR research. J Cardiovasc Magn Reson 2014 Jan 16;16(Supplement 1):P89. [doi: 10.1186/1532-429x-16-s1-p89]
17. Chen C, Turner SP, Sholle ET, Brown SW, Blau VL, Brouwer JP, et al. Evaluation of a REDCap-based workflow for supporting federal guidance for electronic informed consent. AMIA Jt Summits Transl Sci Proc 2019;2019:163-172 [FREE Full text] [Medline: 31258968]
18. Lee CA, Gamino D, Lore M, Donelson C, Windsor LC. Use of research electronic data capture (REDCap) in a sequential multiple assignment randomized trial (SMART): a practical example of automating double randomization. BMC Med Res Methodol 2023 Jul 06;23(1):162 [FREE Full text] [doi: 10.1186/s12874-023-01986-6] [Medline: 37415099]

19. Doyle S, Pavlos R, Carlson SJ, Barton K, Bhuiyan M, Boeing B, et al. Efficacy of digital health tools for a pediatric patient registry: semistructured interviews and interface usability testing with parents and clinicians. JMIR Form Res 2022 Jan 17;6(1):e29889 [FREE Full text] [doi: 10.2196/29889] [Medline: 35037889]
20. Garcia KK, Abrahão AA. Research development using REDCap software. Healthc Inform Res 2021 Oct;27(4):341-349 [FREE Full text] [doi: 10.4258/hir.2021.27.4.341] [Medline: 34788915]
21. Partners - REDCap. REDCap. URL: https://projectredcap.org/partners/ [accessed 2023-04-21]
22. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
23. 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 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
24. 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. Washington, DC: American Psychological Association; 2012:57-71.
25. Bernard HR. Research Methods in Anthropology: Qualitative and Quantitative Approaches. Lanham, MD: Rowman & Littlefield Publishers; 2018.
26. Bernard HR, Wutich A, Ryan GW. Analyzing Qualitative Data: Systematic Approaches. Thousand Oaks, CA: SAGE Publications; 2016.
27. Bargas-Avila JA, Brenzikofer O, Roth SP, Tuch AN, Orsini S, Opwis K. Simple but crucial user interfaces in the world wide web: introducing 20 guidelines for usable web form design. In: Mátrai R, editor. User Interfaces. London, UK: IntechOpen; May 1, 2010.
28. Conrad FG, Schober MF. Clarifying survey questions when respondents don't know they need clarification. National Center for Education Statistics. 2001. URL: https://nces.ed.gov/FCSM/pdf/2001FCSM_Conrad.pdf [accessed 2023-04-24]
29. ClinCapture homepage. ClinCapture. URL: https://www.clincapture.com/ [accessed 2023-05-09]
30. Pawelek J, Baca-Motes K, Pandit JA, Berk BB, Ramos E. The power of patient engagement with electronic health records as research participants. JMIR Med Inform 2022 Jul 08;10(7):e39145 [FREE Full text] [doi: 10.2196/39145] [Medline: 35802410]
31. SurveyMonkey homepage. SurveyMonkey. URL: https://www.surveymonkey.com/ [accessed 2023-05-09]
32. Qualtrics XM: the leading experience management software. Qualtrics XM. URL: https://www.qualtrics.com/ [accessed 2023-05-09]
33. Soni H, Ivanova J, Wilczewski H, Bailey A, Ong T, Narma A, et al. Virtual conversational agents versus online forms: patient experience and preferences for health data collection. Front Digit Health 2022 Oct 13;4:954069 [FREE Full text] [doi: 10.3389/fdgth.2022.954069] [Medline: 36310920]
34. Ponathil A, Ozkan F, Welch B, Bertrand J, Chalil Madathil K. Family health history collected by virtual conversational agents: an empirical study to investigate the efficacy of this approach. J Genet Couns 2020 Dec 03;29(6):1081-1092. [doi: 10.1002/jgc4.1239] [Medline: 32125052]
35. Heimlich R. More use cell phones to get health information. Pew Research Center. 2012 Nov 14. URL: https://www.pewresearch.org/fact-tank/2012/11/14/more-use-cell-phones-to-get-health-information/ [accessed 2022-04-18]

Abbreviations

EDC: electronic data capture
EHR: electronic health record
HIPAA: Health Insurance Portability and Accountability Act
REDCap: Research Electronic Data Capture

Edited by C Lovis; submitted 09.06.23; peer-reviewed by S Wang, C Chen; comments to author 12.03.24; revised version received 10.04.24; accepted 04.05.24; published 25.06.24.

Please cite as:
Soni H, Ivanova J, Wilczewski H, Ong T, Ross JN, Bailey A, Cummins M, Barrera J, Bunnell B, Welch B
User Preferences and Needs for Health Data Collection Using Research Electronic Data Capture: Survey Study
JMIR Med Inform 2024;12:e49785
URL: https://medinform.jmir.org/2024/1/e49785
doi:10.2196/49785
PMID:

Generalization of a Deep Learning Model for Continuous Glucose Monitoring–Based Hypoglycemia Prediction: Algorithm Development and Validation Study

Jian Shao[1], PhD; Ying Pan[2], PhD; Wei-Bin Kou[3], PhD; Huyi Feng[4], PhD; Yu Zhao[1], BBA; Kaixin Zhou[1], PhD; Shao Zhong[2], PhD

1

2

3

4

Corresponding Author:
Shao Zhong, PhD

Abstract

Background: Predicting hypoglycemia while maintaining a low false alarm rate is a challenge for the wide adoption of continuous glucose monitoring (CGM) devices in diabetes management. One small study suggested that a deep learning model based on the long short-term memory (LSTM) network had better performance in hypoglycemia prediction than traditional machine learning algorithms in European patients with type 1 diabetes. However, given that many well-recognized deep learning models perform poorly outside the training setting, it remains unclear whether the LSTM model could be generalized to different populations or patients with other diabetes subtypes.

Objective: The aim of this study was to validate LSTM hypoglycemia prediction models in more diverse populations and across a wide spectrum of patients with different subtypes of diabetes.

Methods: We assembled two large data sets of patients with type 1 and type 2 diabetes. The primary data set including CGM data from 192 Chinese patients with diabetes was used to develop the LSTM, support vector machine (SVM), and random forest (RF) models for hypoglycemia prediction with a prediction horizon of 30 minutes. Hypoglycemia was categorized into mild (glucose=54-70 mg/dL) and severe (glucose<54 mg/dL) levels. The validation data set of 427 patients of European-American ancestry in the United States was used to validate the models and examine their generalizations. The predictive performance of the models was evaluated according to the sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).

Results: For the difficult-to-predict mild hypoglycemia events, the LSTM model consistently achieved AUC values greater than 97% in the primary data set, with a less than 3% AUC reduction in the validation data set, indicating that the model was robust and generalizable across populations. AUC values above 93% were also achieved when the LSTM model was applied to both type 1 and type 2 diabetes in the validation data set, further strengthening the generalizability of the model. Under different satisfactory levels of sensitivity for mild and severe hypoglycemia prediction, the LSTM model achieved higher specificity than the SVM and RF models, thereby reducing false alarms.

Conclusions: Our results demonstrate that the LSTM model is robust for hypoglycemia prediction and is generalizable across populations or diabetes subtypes. Given its additional advantage of false-alarm reduction, the LSTM model is a strong candidate to be widely implemented in future CGM devices for hypoglycemia prediction.

(JMIR Med Inform 2024;12:e56909) doi:10.2196/56909

KEYWORDS

hypoglycemia prediction; hypoglycemia; hypoglycemic; blood sugar; prediction; predictive; deep learning; generalization; machine learning; glucose; diabetes; continuous glucose monitoring; type 1 diabetes; type 2 diabetes; LSTM; long short-term memory

Introduction

Diabetes is a serious long-term disease with considerable influence on global health [1]. Type 1 diabetes mellitus (T1DM) is a disease in which the pancreas produces little or no insulin [2], whereas insulin resistance and insufficient insulin are the primary contributors to the development of type 2 diabetes mellitus (T2DM) [3]. Although the pathogenic mechanisms of

T1DM and T2DM are different, glucose-lowering treatments such as insulin administration are the common leading cause of hypoglycemia events in patients with both diabetes subtypes [4]. Severe hypoglycemia is a frequent phenomenon in patients with T1DM, with an annual prevalence of 30%-40% [5]. Although the risk of severe hypoglycemia in patients with T2DM is relatively lower, 46%-58% of these patients were reported to have experienced mild hypoglycemia symptoms over a 6-month period [6]. Patients experiencing frequent hypoglycemia events have 1.5-6.0 times increased risks of cardiovascular events and mortality than those without such events [7]. Patients with T2DM from Southeast Asia appear to have an elevated risk of hypoglycemia, as these patients are more often treated with a premixed insulin formulation, are younger, and have a lower BMI than those of their counterparts from Western countries [8-11]. Given that demographic and clinical factors such as ethnic group, diabetes subtype, and BMI are all important components of the complex risk profile of hypoglycemia, accurate risk prediction and prevention of hypoglycemia across populations and diabetes types remain significant challenges in diabetes management.

Recently, continuous glucose monitoring (CGM) has demonstrated good potential to predict hypoglycemia. For patients who wear insulin pumps or those who require multiple daily insulin injections, hypoglycemia prediction based on CGM data could provide a timely warning of impending hypoglycemia for the individual to take immediate action and increase their glucose levels. CGM devices are designed to produce time-series data by recording interstitial glucose concentrations within a relatively short interval of 5-15 minutes over a few days. Therefore, it is possible to leverage the early glucose readings to predict hypoglycemia events over the short-to-medium time horizon. Time-series forecast algorithms such as autoregressive and moving-average algorithms were first adopted to utilize the short-term temporal features of CGM data to predict hypoglycemia [12-15]. A small study including 17 patients with T1DM showed that these CGM-based algorithms achieved 86% sensitivity but only 58% specificity in hypoglycemia prediction [16]. Similar results from studies implementing these time-series forecast algorithms indicated that the low specificity might frequently generate false alarms, leading to discontinuation of CGM use in hypoglycemia prevention [17,18].

To improve the sensitivity and particularly the specificity of hypoglycemia prediction, both traditional machine learning algorithms such as support vector machine (SVM) and random forest (RF) models, along with deep learning models such as the convolutional neural network and long short-term memory network (LSTM) have been used to leverage more temporal features of CGM data [19-25]. When the features, including the mean of glucose and range of time in hyperglycemia, based on CGM data collected over the previous 6 hours were fed into the RF model, hypoglycemia prediction achieved a sensitivity of 93% and a specificity of 91% in a study of 112 patients with T1DM [26]. More recently, when an LSTM deep learning model was implemented on CGM data for hypoglycemia prediction, it achieved a sensitivity of 97% with remarkably few false alarms (0.9 false alarms per week) on a test data set including 10 patients with T1DM, thereby illuminating a path toward the widespread clinical adoption of CGM in hypoglycemia prediction [27].

However, a well-known challenge in implementing predictive models is their generalization [28]. The predictive performance of models could be substantially reduced when used in a setting that is not well-represented by the training data set [29,30]. This is particularly relevant in the case of hypoglycemia prediction, as the previously developed models for this purpose were mostly trained on a small data set of patients with T1DM from Western populations. In addition, the lack of a common test data set rendered the comparison of predictive performances between models unreliable. With recent improvements in measurement accuracy, CGM devices have also gained momentum and have begun to be adopted more widely for the management of T2DM, including in developing countries. Therefore, the established hypoglycemia prediction models should be validated in more diverse populations and over a wide spectrum of patients with different types of diabetes.

We hypothesized that the promising LSTM model for hypoglycemia prediction from CGM data could maintain good predictive performance in different settings for different populations. In this study, we assembled two large CGM data sets from China and the United States, both including patients with T1DM and patients with T2DM. We developed the LSTM model on the Chinese data set and then examined the model performance in the data set from European-Americans in the United States. Apart from exploring the model's generalization ability for T1DM and T2DM separately, we also compared the predictive performance of the LSTM model with that of SVM and RF models to further indicate its translational potential.

Methods

Ethical Considerations

The study protocol was approved by the ethics committees of Kunshan Hospital Affiliated to Jiangsu University (2023-03-014-H01-K01) and the study was performed in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from each participant before taking the measurements. The data analyzed were anonymized. All participants volunteered to participate in the project with no compensation provided.

Data Collection

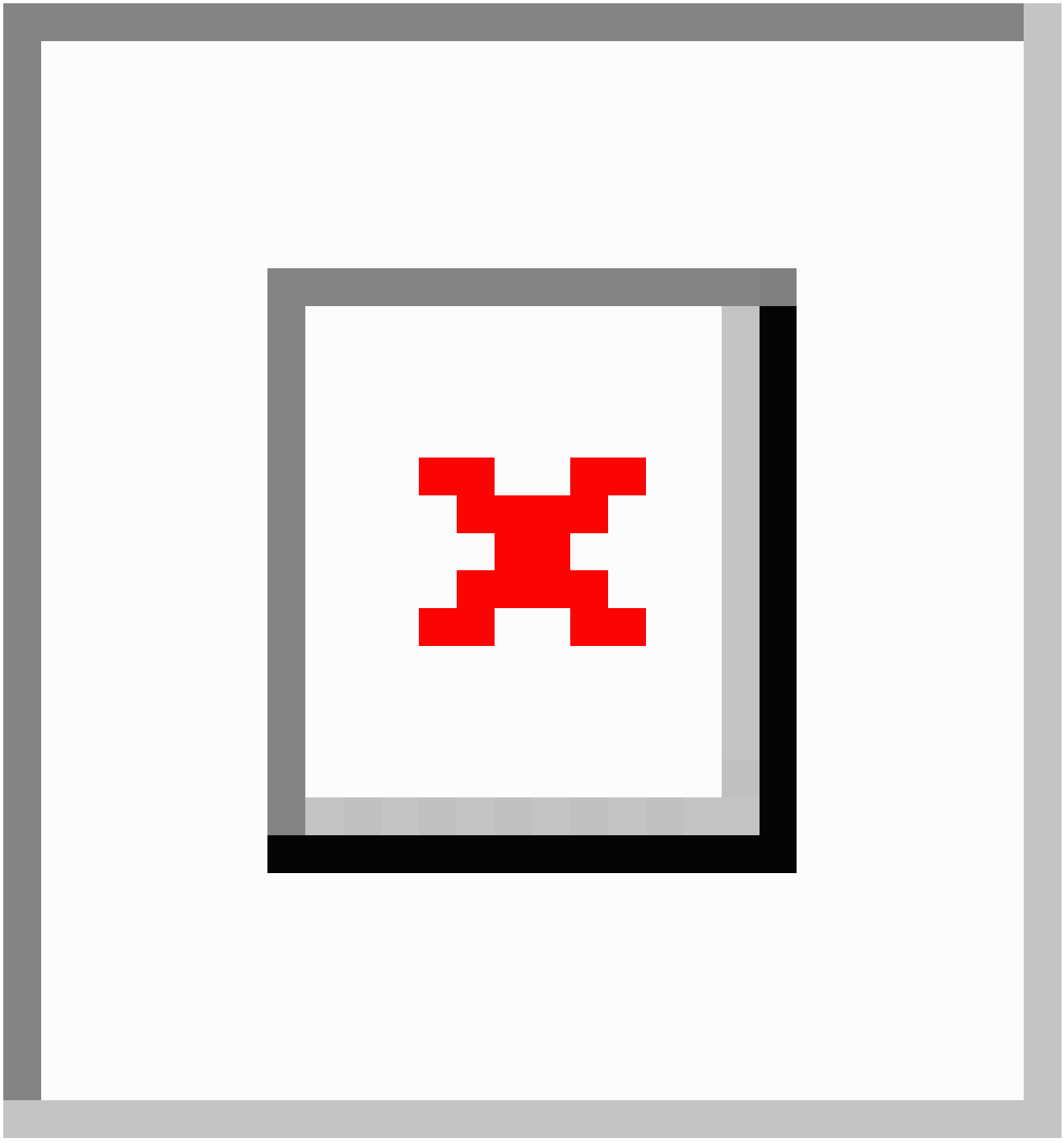
We collected a primary data set comprising 1578 days of CGM data collected from 264 Chinese people with diabetes to develop a deep learning model for hypoglycemia prediction. The individuals' glucose levels were monitored using the Medtronic MiniMed CGM device, which requires calibration according to self-monitored blood glucose levels. This CGM device can record glucose levels every 5 minutes over 3 days.

The mean absolute relative difference (MARD) was used to evaluate the quality of the CGM data. The MARD represents the average of the absolute error between all CGM values and matched reference values. A small MARD indicates that the CGM readings are close to the reference glucose value, whereas a larger MARD percentage indicates greater discrepancies between the CGM and reference glucose values. Each individual

had at least 5 self-monitoring of blood glucose (SMBG) measurements. As reference glucose values, the SMBG was used to calculate the MARD of CGM data. The data for 72 participants were filtered out because their MARD was higher than 15%, leaving data for 192 participants with 808 days of CGM data for analysis.

To examine whether the deep learning model trained and developed with data from the Chinese population could be generalized to a different population, we assembled a large validation data set that mainly comprised data from individuals of European-American ancestry. The validation data set shared by the A1c-Derived Average Glucose study group includes 507 participants and 7299 days of CGM data, also collected with Medtronic MiniMed devices [31]. After filtering out individuals without diabetes, 427 patients with either T1DM or T2DM were included to validate the model. This validation data set was split into two groups: the T1DM group of 268 participants with 3932 days of CGM data and the T2DM group of 159 participants with 2259 days of CGM data. Figure 1 provides the flowchart of exclusion criteria for the primary data set and validation data set.

Figure 1. Flowchart of exclusion criteria for the primary data set and validation data set. MARD: mean absolute relative difference.



Outcome

The glucose values reported by CGM devices were classified into three categories: nonhypoglycemic level (glucose>70 mg/dL), mild hypoglycemic level (glucose=54-70 mg/dL), and severe hypoglycemic level (glucose<54 mg/dL) according to the international consensus on CGM utility [32].

Data Preprocessing

The primary data set consisting of 192 patients was randomly split into three disjoint data sets, namely the training data set, development data set, and test data set, at a 7:1.5:1.5 ratio. The training data set was used to train the model, whereas the development data set was used to select the hyperparameters in the training process. The test data set was used to evaluate the performance of the developed model.

The CGM sensor may fail to detect a valid glucose level, resulting in the CGM device missing glucose values continuously. To preserve as much of the CGM data as possible, we divided an individual's CGM data into different segments at the time points of missing data rather than discarding all of the CGM data. A segment was removed if it was shorter than 6 hours (72 data points). We set each glucose value reported by the CGM device as a predictive target if there were sufficient data prior to the target time at which the predictive target was located. The data used to predict the hypoglycemic level of the predictive target were retrieved from a 6-hour time window spanning from −390 minutes to −30 minutes of the target time. After preprocessing the primary data set, the training, development, and test data sets included 100,879, 21,895, and 21,324 samples generated from 134, 29, and 29 participants, respectively. Similarly, the T1DM group and T2DM group from the validation data set contained 712,018 and 405,224 samples generated from 268 and 159 participants, respectively.

Model Development

We used the common bidirectional LSTM model containing both forward and backward layers to capture the long-range temporal features in the time-series CGM data and to combine these features with context factors [33]. Each LSTM layer consists of 128 memory cells [34]. We chose a set of context factors, including gender, age, diabetes type, and hemoglobin A_{1c} value, to capture the background risk of hypoglycemia and enhance the model's predictive performance [26]. Therefore, each input data sample included 72 points of CGM data collected during 6 hours and the context factors. The output was the probability of the target glucose value being at the nonhypoglycemic level, mild hypoglycemic level, and severe hypoglycemic level.

We trained the LSTM model to predict the categories of a CGM value within 30 minutes on the prediction horizon. The training process would be terminated if the accuracy failed to increase for 10 consecutive epochs. We used root mean square propagation [35] as the optimizer and set the mini batch size to 64. The LSTM model was developed using the Python package Keras [36]. We also developed models to implement the SVM and RF algorithms for comparison. The SVM model was developed using the radial basis function as the kernel function, which was also used in previous studies of hypoglycemia prediction [37]. The RF model included 100 trees and was developed with the Scikit-learn Python package [38] under default parameters. The input to the SVM and RF models was the same as that used for the LSTM model.

Model Evaluation

Sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) were used to evaluate model performance. The label for each sample was the category of a single CGM data point. Sensitivity and specificity indicate the proportion of the labels of CGM data points that were correctly predicted. The DeLong method was used to measure the 95% CIs for the AUC values [39]. All methods of evaluation were developed using Python and the pROC R package [40].

Results

Characteristics of the Data Set

Table 1 summarizes the characteristics of the primary data set and the validation data set. As expected, the average age of patients with T1DM was lower than that of the patients with T2DM in both data sets (Wilcoxon rank sum test, $P<.001$).

Table . Characteristics of the primary data set and validation data set.

Variables		Primary data set		Validation data set	
		Type 2 diabetes (n=175)	Type 1 diabetes (n=17)	Type 2 diabetes (n=159)	Type 1 diabetes (n=268)
Age (years), mean (SD)		53.30 (11.78)	40.59 (13.02)	55.64 (9.32)	43.06 (12.85)
Women, n (%)		51 (29.14)	11 (64.71)	81 (50.94)	140 (52.24)
Predictive targets, n					
	Nonhypoglycemia	129,609	12,029	396,415	660,111
	Mild hypoglycemia	1350	336	5985	28,287
	Severe hypoglycemia	608	166	2824	23,620
Hemoglobin A_{1c} (%), mean (SD)		7.69 (1.71)	8.46 (2.22)	7.01 (1.24)	7.51 (1.30)

Model Performance on the Primary Test Data Set

Using the primary data set from 192 individuals, the three models of LSTM, SVM, and RF were trained and we then evaluated their performance based on the AUC. At the mild hypoglycemic level, the LSTM model achieved an AUC of 97.22% (95% CI 96.78%-97.66%), which was significantly higher than the AUC of 94.33% (95% CI 93.13%-95.53%) and 94.81% (95% CI 93.72%-95.91%) achieved by the SVM and RF models, respectively (both $P<.001$). At the severe hypoglycemic level, the LSTM model achieved an AUC of 99.64% (95% CI 99.53%-99.76%), which was significantly higher than the AUC of 98.30% (95% CI 98.00%-98.60%) and 97.88% (95% CI 96.93%-98.83%) achieved by the SVM and RF models, respectively (both $P<.001$). These results demonstrated that the LSTM model could outperform the SVM and RF models in predicting hypoglycemia.

Model Generalization on the Validation Data Set

We then utilized the validation data set from 427 European-Americans to evaluate the generalization of the LSTM model developed from our primary data set of 192 Chinese individuals. The LSTM model achieved an AUC of 94.61% (95% CI 94.51%-94.71%) for mild hypoglycemia, which was significantly higher than the AUC of 92.59% (95% CI 92.48%-92.71%) and 91.43% (95% CI 91.28%-91.58%) achieved by the SVM and RF models, respectively (both $P<.001$). The LSTM model achieved an AUC of 96.40% (95% CI 96.25%-96.55%) for severe hypoglycemia, which was significantly higher than the AUC of 95.27% (95% CI 95.15%-95.39%) and 95.17% (95% CI 95.01%-95.32%) achieved by SVM and RF models, respectively (both $P<.001$). Although AUC values of the LSTM model decreased by approximately 3% in the validation data set compared to those from the primary test data set, the overall AUC was still higher than 94%, indicating that the LSTM model could accurately predict hypoglycemia in a different population.

Next, the generalizability of the LSTM model to various disease subtypes was evaluated in the subgroups of T1DM and T2DM from the validation data set. For T1DM, the LSTM model achieved an AUC of 93.49% (95% CI 93.38%-93.61%) at the mild hypoglycemia level, which was significantly higher than the AUC of 90.92% (95% CI 90.78%-91.06%) and 89.74% (95% CI 89.57%-89.92%) achieved by the SVM and RF models, respectively (both $P<.001$). In addition, the LSTM model achieved an AUC of 95.89% (95% CI 95.73%-96.05%) at the severe hypoglycemia level, which was significantly higher than the AUC of 94.06% (95% CI 93.91%-94.21%) and 94.53% (95% CI 94.37%-94.70%) achieved by the SVM and RF models, respectively (both $P<.001$).

For T2DM, the LSTM model achieved an AUC of 96.83% (95% CI 96.66%-97.01%) at the mild hypoglycemia level, which was significantly higher than the AUC of 95.72% (95% CI 95.51%-95.93%) and 94.08% (95% CI 93.73%-94.43%) achieved by the SVM and RF models, respectively (both $P<.001$). In addition, the LSTM model achieved an AUC of 97.65% (95% CI 97.27%-98.04%) at the severe hypoglycemia level, which was significantly higher than the AUC of 96.02% (95% CI 95.70%-96.34%) and 95.71% (95% CI 95.23%-96.19%) achieved by the SVM and RF models, respectively (both $P<.001$).

The AUCs of the LSTM model were consistently higher than those from the SVM and RF models in both the T1DM and T2DM data sets. Taken together, these results demonstrated that the LSTM model could be generalized to different diabetes subtypes without significant loss of predictive performance.

Comparison of the False Alarm Rate

Finally, we examined whether the LSTM model could achieve a low false alarm rate (ie, high specificity) under satisfactory sensitivity. According to previous studies of hypoglycemia prediction, we set the model parameters to fix the satisfactory sensitivity level at 90% and 95% for mild and severe hypoglycemia prediction, respectively [21,26,37]. As shown in Table 2, while maintaining a sensitivity of 90% for mild hypoglycemia, which is difficult to predict, the LSTM model could achieve a specificity of 88.43%, which was higher than the specificity obtained from the SVM and RF models. For severe hypoglycemia, when a higher satisfactory sensitivity rate of 95% was set, the LSTM model achieved a specificity of 87.34%, which was higher than that obtained from the SVM model. Moreover, the RF model could not achieve a sensitivity of 95% for the severe hypoglycemic level. Taken together, these results demonstrated that the LSTM model could maintain a lower false alarm rate than the SVM and RF models in clinically practical settings.

Table . Specificity and sensitivity of the three models on the validation data set.

	Mild hypoglycemic level		Severe hypoglycemic level	
	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)
LSTM[a]	88.43	90.00	87.34	95.00
SVM[b]	82.57	90.00	80.67	95.00
RF[c]	82.65	90.00	Not determined	Not achieved

[a]LSTM: long short-term memory.

[b]SVM: support vector machine.

[c]RF: random forest.

Discussion

Principal Findings

In this study, we assembled two large CGM data sets from China and the United States to develop and validate an LSTM deep learning model for hypoglycemia prediction. The LSTM model could maintain good predictive performance when applied to data sets from a different ethnic population or any common subtype of diabetes. The LSTM model could also predict both mild and severe hypoglycemia with higher accuracy than the traditional SVM and RF models. While targeting clinically meaningful high sensitivity, the LSTM model could achieve high specificity, thereby reducing the rate of false alarms.

Compared with the models tested without external validation in most previous studies of hypoglycemia prediction, we developed an LSTM model and validated the model in a data set from a different population to examine its generalizability [27]. There are considerable differences in dietary structure and clinical practice between China and the United States, which are among the many factors that might affect the risk of hypoglycemia. Previous studies demonstrated that clinical models trained in one population could result in an AUC reduction as great as 15% when applied to a distinct population [41-43]. However, the LSTM model derived from our Chinese training data set maintained high prediction performance (AUC>93%) with only a minor loss of 3% in the US data set, indicating good generalizability of the model. As CGM devices are becoming more widely adopted, the generalizability of the LSTM model could be further improved by training the model with data from multiple populations or can be fine-tuned for the target population using a transfer-learning approach [44].

We also examined the generalizability of the LSTM model on another dimension of diabetes pathogenicity. Given the different pathogenic mechanisms between T1DM and T2DM, hypoglycemia occurring in different diabetes subtypes would be expected to be preceded by various patterns of glucose fluctuation, which could be leveraged by the LSTM model for prediction. Therefore, the model was expected to lose predictive performance when the training and validation data sets had different proportions of diabetes subtypes. Indeed, we observed a higher AUC value for T2DM than for T1DM in the validation data set, which was likely due to the fact that our training data set primarily consisted of individuals with T2DM. However, for either subtype of diabetes, the LSTM model consistently maintained an AUC value above 93%, indicating the good generalizability of the model. With the increasing popularity of CGM usage in the management of all subtypes of diabetes, the LSTM model could be further improved by using larger training data sets with a wider representation of the various diabetes subtypes.

Achieving high sensitivity has been the main focus of previous models for hypoglycemia prediction, as severe hypoglycemia requires immediate external intervention [15,32]. With the sacrifice of high specificity, false alarms became an obstacle for the safe and widespread use of CGM devices [45-47]. False-alarm fatigue could lead to users ignoring the true alarms of hypoglycemia and contribute to the discontinuation of CGM use [45]. Moreover, glucose control could be compromised, as CGM users may frequently take action to elevate their glucose level when a false alarm is generated [46]. Therefore, it is imperative to balance the false alarm rate with sufficient sensitivity of the prediction. In this study, we demonstrated that the LSTM model would generate fewer false alarms than the traditional machine learning models under satisfactory sensitivity rates of 90% and 95% for mild and severe hypoglycemia, respectively. Therefore, the balanced hypoglycemia prediction performance from the LSTM model demonstrated that it has potential to promote the use of CGM in a variety of clinical settings.

One reason for the better predictive performance of the LSTM model than the SVM and RF models might be that the LSTM algorithm is more suitable for analyzing sequential data. CGM data are a type of sequential data that are generated in time order. The LSTM algorithm consists of memory cells that learn the sequential nature of observations within CGM data [48]. The input of one memory cell is the glucose value taken at one time point and then the LSTM takes all of the glucose values as inputs sequentially. Every memory cell retains the relevant information and discards irrelevant information for the predictive task, and then the relevant information in one cell is delivered to the next cell [49-53]. With this sequential structure, LSTM networks incorporate CGM data from the past to accurately make predictions of hypoglycemia risk in the near future.

Limitations

There are several limitations of this study. Although we tested the generalizability of the LSTM model using two data sets from China and the United States, further validation might still be required for application of the model in other countries. Similarly, as only T1DM and T2DM were included in our data sets, the model should be tested with wider and more representative training data sets to validate its utility on other minority subtypes of diabetes. Moreover, data from only one CGM device manufacturer were available for this study. Thus, it is unknown whether the model would perform equally well with data collected from other devices such as factory-calibrated CGM or noninvasive CGM devices. However, given that all of the devices were strictly calibrated by finger-stick glucose values, the fluctuation patterns and temporal dependence of CGM data, which are key factors for the LSTM prediction task, should be largely captured by any certified CGM device. Moreover, the performance of the LSTM model for hypoglycemia prediction will need to be further validated in a CGM data set without missing data.

Conclusions

We developed an accurate LSTM model for mild and severe hypoglycemia prediction using a large data set of 619 patients with diabetes from China and the United States. The model could be robustly generalized to different populations or any common subtype of diabetes. Moreover, while maintaining satisfactory levels of sensitivity, the model could also achieve high specificity, indicating its potential to mitigate the hypoglycemia false-alarm fatigue that is frequently observed in clinical practice. Taken together, we demonstrated that the

LSTM model is a strong candidate algorithm to be further tested and implemented for the wider clinical adoption of CGM.

Acknowledgments

We thank all of the involved clinicians and researchers for data collection and assistance. This study was funded by the National Key R&D Program of China (SQ2022YFB3200174) and Suzhou Science and Technology Project (SKY2022025).

Data Availability

Requests for access to the study data should be directed to the corresponding author.

Conflicts of Interest

None declared.

References

1. Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. Phys Ther 2008 Nov;88(11):1254-1264. [doi: 10.2522/ptj.20080020] [Medline: 18801858]
2. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. Lancet 2014 Jan;383(9911):69-82. [doi: 10.1016/S0140-6736(13)60591-7] [Medline: 25130995]
3. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. Lancet 2017 Jun 3;389(10085):2239-2251. [doi: 10.1016/S0140-6736(17)30058-2] [Medline: 28190580]
4. Cryer PE. The barrier of hypoglycemia in diabetes. Diabetes 2008 Dec;57(12):3169-3176. [doi: 10.2337/db08-1084] [Medline: 19033403]
5. Frier BM. The incidence and impact of hypoglycemia in type 1 and type 2 diabetes. Inter Diab Monitor 2009;21(6):210-218.
6. Silbert R, Salcido-Montenegro A, Rodriguez-Gutierrez R, Katabi A, McCoy RG. Hypoglycemia among patients with type 2 diabetes: epidemiology, risk factors, and prevention strategies. Curr Diab Rep 2018 Jun 21;18(8):53. [doi: 10.1007/s11892-018-1018-0] [Medline: 29931579]
7. International Hypoglycaemia Study Group. Hypoglycaemia, cardiovascular disease, and mortality in diabetes: epidemiology, pathogenesis, and management. Lancet Diabetes Endocrinol 2019 May;7(5):385-396. [doi: 10.1016/S2213-8587(18)30315-2] [Medline: 30926258]
8. Chan JCN, Malik V, Jia W, et al. Diabetes in Asia: epidemiology, risk factors, and pathophysiology. JAMA 2009 May 27;301(20):2129-2140. [doi: 10.1001/jama.2009.726] [Medline: 19470990]
9. Kalra S, Balhara YPS, Sahay BK, Ganapathy B, Das AK. Why is premixed insulin the preferred insulin? Novel answers to a decade-old question. J Assoc Physicians India 2013 Jan;61(1 Suppl):9-11. [Medline: 24482980]
10. Goh SY, Hussein Z, Rudijanto A. Review of insulin-associated hypoglycemia and its impact on the management of diabetes in Southeast Asian countries. J Diabetes Investig 2017 Sep;8(5):635-645. [doi: 10.1111/jdi.12647] [Medline: 28236664]
11. Aschner P, Sethi B, Gomez-Peralta F, et al. Insulin glargine compared with premixed insulin for management of insulin-naïve type 2 diabetes patients uncontrolled on oral antidiabetic drugs: the open-label, randomized GALAPAGOS study. J Diabetes Complications 2015 Aug;29(6):838-845. [doi: 10.1016/j.jdiacomp.2015.04.003] [Medline: 25981123]
12. Eren-Oruklu M, Cinar A, Quinn L, Smith D. Estimation of future glucose concentrations with subject-specific recursive linear models. Diabetes Technol Ther 2009 Apr;11(4):243-253. [doi: 10.1089/dia.2008.0065] [Medline: 19344199]
13. Yang J, Li L, Shi Y, Xie X. An ARIMA model with adaptive orders for predicting blood glucose concentrations and Hypoglycemia. IEEE J Biomed Health Inform 2019 May;23(3):1251-1260. [doi: 10.1109/JBHI.2018.2840690] [Medline: 29993728]
14. Eren-Oruklu M, Cinar A, Rollins DK, Quinn L. Adaptive system identification for estimating future glucose concentrations and hypoglycemia alarms. Automatica (Oxf) 2012 Aug;48(8):1892-1897. [doi: 10.1016/j.automatica.2012.05.076] [Medline: 22865931]
15. Dassau E, Cameron F, Lee H, et al. Real-time hypoglycemia prediction suite using continuous glucose monitoring: a safety net for the artificial pancreas. Diabetes Care 2010 Jun;33(6):1249-1254. [doi: 10.2337/dc09-1487] [Medline: 20508231]
16. Bayrak ES, Turksoy K, Cinar A, Quinn L, Littlejohn E, Rollins D. Hypoglycemia early alarm systems based on recursive autoregressive partial least squares models. J Diabetes Sci Technol 2013 Jan 1;7(1):206-214. [doi: 10.1177/193229681300700126] [Medline: 23439179]
17. Tansey M, Laffel L, Cheng J, et al. Satisfaction with continuous glucose monitoring in adults and youths with type 1 diabetes. Diabet Med 2011 Sep;28(9):1118-1122. [doi: 10.1111/j.1464-5491.2011.03368.x] [Medline: 21692844]
18. Ramchandani N, Arya S, Ten S, Bhandari S. Real-life utilization of real-time continuous glucose monitoring: the complete picture. J Diabetes Sci Technol 2011 Jul 1;5(4):860-870. [doi: 10.1177/193229681100500407] [Medline: 21880227]

19. Georga EI, Protopappas VC, Ardigò D, Polyzos D, Fotiadis DI. A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions. Diabetes Technol Ther 2013 Aug;15(8):634-643. [doi: 10.1089/dia.2012.0285] [Medline: 23848178]
20. 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]
21. Mosquera-Lopez C, Dodier R, Tyler NS, 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. [doi: 10.1089/dia.2019.0458] [Medline: 32297795]
22. Gu W, Zhou Z, Zhou Y, He M, Zou H, Zhang L. Predicting blood glucose dynamics with multi-time-series deep learning. Presented at: SenSys '17: 15th ACM Conference on Embedded Network Sensor Systems; Nov 5 to 8, 2017; Delft, The Netherlands. [doi: 10.1145/3131672.3136965]
23. Chen J, Li K, Herrero P, Zhu T, Georgiou P. Dilated recurrent neural network for short-time prediction of glucose concentration. Presented at: 3rd International Workshop on Knowledge Discovery in Healthcare Data co-located with the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018); Jul 13, 2018; Stockholm, Sweden. [doi: 10.1007/s41666-020-00068-2]
24. Doike T, Hayashi K, Arata S, Mohammad KN, Kobayashi A, Niitsu K. A blood glucose level prediction system using machine learning based on recurrent neural network for Hypoglycemia prevention. Presented at: 2018 16th IEEE International New Circuits and Systems Conference (NEWCAS); Jun 24 to 27, 2018; Montreal, QC. [doi: 10.1109/NEWCAS.2018.8585468]
25. Li J, Ma X, Tobore I, et al. A novel CGM metric-gradient and combining mean sensor glucose enable to improve the prediction of nocturnal hypoglycemic events in patients with diabetes. J Diabetes Res 2020 Nov;2020:8830774. [doi: 10.1155/2020/8830774] [Medline: 33204733]
26. Dave D, DeSalvo DJ, Haridas B, et al. Feature-based machine learning model for real-time hypoglycemia prediction. J Diabetes Sci Technol 2021 Jul;15(4):842-855. [doi: 10.1177/1932296820922622] [Medline: 32476492]
27. Mosquera-Lopez C, Dodier R, Tyler N, Resalat N, Jacobs P. Leveraging a big dataset to develop a recurrent neural network to predict adverse glycemic events in type 1 diabetes. IEEE J Biomed Health Inform 2019 Apr 17. [doi: 10.1109/JBHI.2019.2911701] [Medline: 30998484]
28. Zhang Y, Wu H, Liu H, Tong L, Wang MD. Improve model generalization and robustness to dataset bias with bias-regularized learning and domain-guided augmentation. arXiv. Preprint posted online on Oct 12, 2019. [doi: 10.48550/arXiv.1910.06745]
29. Kortylewski A, Egger B, Schneider A, Gerig T, Morel-Forster A, Vetter T. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Jun 16 to 17, 2019; Long Beach, CA, USA URL: https://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8972688 [accessed 2024-05-24] [doi: 10.1109/CVPRW.2019.00279]
30. Tian Y, Chen W, Zhou T, Li J, Ding K, Li J. Establishment and evaluation of a multicenter collaborative prediction model construction framework supporting model generalization and continuous improvement: a pilot study. Int J Med Inform 2020 Sep;141:104173. [doi: 10.1016/j.ijmedinf.2020.104173] [Medline: 32531725]
31. Nathan DM, Kuenen J, Borg R, et al. Translating the A1C assay into estimated average glucose values. Diabetes Care 2008 Aug;31(8):1473-1478. [doi: 10.2337/dc08-0545] [Medline: 18540046]
32. Danne T, Nimri R, Battelino T, et al. International consensus on use of continuous glucose monitoring. Diabetes Care 2017 Dec;40(12):1631-1640. [doi: 10.2337/dc17-1600] [Medline: 29162583]
33. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. Neural Comput 2000 Oct;12(10):2451-2471. [doi: 10.1162/089976600300015015] [Medline: 11032042]
34. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
35. Hinton G, Srivastava N, Swersky K. Neural networks for machine learning. Lecture 6a. Overview of mini-batch gradient descent. Computer Science University of Toronto. URL: http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf [accessed 2024-05-15]
36. Keras. URL: https://keras.io/ [accessed 2024-05-13]
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. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011 Nov 1;12:2825-2830. [doi: 10.5555/1953048.2078195]
39. 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. [doi: 10.2307/2531595] [Medline: 3203132]
40. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011 Mar 17;12:77. [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]

41. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 1993 Nov 24;270(20):2478-2486. [doi: 10.1001/jama.1993.03510200084037] [Medline: 8230626]
42. Adrie C, Francais A, Alvarez-Gonzalez A, et al. Model for predicting short-term mortality of severe sepsis. Crit Care 2009 May;13(3):R72. [doi: 10.1186/cc7881] [Medline: 19454002]
43. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016 Jun 22;353:i3140. [doi: 10.1136/bmj.i3140] [Medline: 27334381]
44. Torrey L, Shavlik J. Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: IGI Global; 2010:242-264. [doi: 10.4018/978-1-60566-766-9]
45. Shivers JP, Mackowiak L, Anhalt H, Zisser H. “Turn it off!”: diabetes device alarm fatigue considerations for the present and the future. J Diabetes Sci Technol 2013 May 1;7(3):789-794. [doi: 10.1177/193229681300700324] [Medline: 23759412]
46. Cryer PE. Glycemic goals in diabetes: trade-off between glycemic control and iatrogenic hypoglycemia. Diabetes 2014 Jul;63(7):2188-2195. [doi: 10.2337/db14-0059] [Medline: 24962915]
47. Wong JC, Foster NC, Maahs DM, et al. Real-time continuous glucose monitoring among participants in the T1D Exchange clinic registry. Diabetes Care 2014 Oct;37(10):2702-2709. [doi: 10.2337/dc14-0303] [Medline: 25011947]
48. Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans Smart Grid 2017 Sep 18;10(1):841-851. [doi: 10.1109/TSG.2017.2753802]
49. Xu Z, Li S, Deng W. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. Presented at: ACPR 2015: 3rd IAPR Asian Conference on Pattern Recognition; Nov 3, 2015; Kuala Lumpur, Malaysia. [doi: 10.1109/ACPR.2015.7486482]
50. Shi X, Jin Y, Dou Q, Heng PA. LRTD: long-range temporal dependency based active learning for surgical workflow recognition. Int J Comput Assist Radiol Surg 2020 Sep;15(9):1573-1584. [doi: 10.1007/s11548-020-02198-9] [Medline: 32588246]
51. Liao J, Liu L, Duan H, et al. Using a convolutional neural network and convolutional long short-term memory to automatically detect aneurysms on 2D digital subtraction angiography images: framework development and validation. JMIR Med Inform 2022 Mar 16;10(3):e28880. [doi: 10.2196/28880] [Medline: 35294371]
52. Athanasiou M, Fragkozidis G, Zarkogianni K, Nikita KS. Long short-term memory–based prediction of the spread of influenza-like illness leveraging surveillance, weather, and Twitter data: model development and validation. J Med Internet Res 2023 Feb 6;25:e42519. [doi: 10.2196/42519] [Medline: 36745490]
53. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, R Niakan Kalhori S. Predicting COVID-19 incidence through analysis of Google trends data in Iran: data mining and deep learning pilot study. JMIR Public Health Surveill 2020 Apr 14;6(2):e18828. [doi: 10.2196/18828] [Medline: 32234709]

Abbreviations

AUC: area under the receiver operating characteristic curve
CGM: continuous glucose monitoring
LSTM: long short-term memory
MARD: mean absolute relative difference
RF: random forest
SMBG: self-monitoring of blood glucose
SVM: support vector machine
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus

Edited by C Lovis; submitted 21.02.24; peer-reviewed by G Lim; revised version received 07.04.24; accepted 04.05.24; published 24.05.24.

Please cite as:
Shao J, Pan Y, Kou WB, Feng H, Zhao Y, Zhou K, Zhong S
Generalization of a Deep Learning Model for Continuous Glucose Monitoring–Based Hypoglycemia Prediction: Algorithm Development and Validation Study
JMIR Med Inform 2024;12:e56909
URL: https://medinform.jmir.org/2024/1/e56909
doi:10.2196/56909

Original Paper

Enhancing Health Equity by Predicting Missed Appointments in Health Care: Machine Learning Study

Yi Yang[1], MSc; Samaneh Madanian[1], PhD; David Parry[2], PhD

[1]Auckland University of Technology, Auckland, New Zealand

[2]Murdoch University, Perth, Australia

Corresponding Author:
Samaneh Madanian, PhD
Auckland University of Technology
6 St Paul Street, AUT WZ Building
Auckland, 1010
New Zealand
Phone: 64 99219999 ext 6539
Email: sam.madanian@aut.ac.nz

Abstract

Background: The phenomenon of patients missing booked appointments without canceling them—known as Did Not Show (DNS), Did Not Attend (DNA), or Failed To Attend (FTA)—has a detrimental effect on patients' health and results in massive health care resource wastage.

Objective: Our objective was to develop machine learning (ML) models and evaluate their performance in predicting the likelihood of DNS for hospital outpatient appointments at the MidCentral District Health Board (MDHB) in New Zealand.

Methods: We sourced 5 years of MDHB outpatient records (a total of 1,080,566 outpatient visits) to build the ML prediction models. We developed 3 ML models using logistic regression, random forest, and Extreme Gradient Boosting (XGBoost). Subsequently, 10-fold cross-validation and hyperparameter tuning were deployed to minimize model bias and boost the algorithms' prediction strength. All models were evaluated against accuracy, sensitivity, specificity, and area under the receiver operating characteristic (AUROC) curve metrics.

Results: Based on 5 years of MDHB data, the best prediction classifier was XGBoost, with an area under the curve (AUC) of 0.92, sensitivity of 0.83, and specificity of 0.85. The patients' DNS history, age, ethnicity, and appointment lead time significantly contributed to DNS prediction. An ML system trained on a large data set can produce useful levels of DNS prediction.

Conclusions: This research is one of the very first published studies that use ML technologies to assist with DNS management in New Zealand. It is a proof of concept and could be used to benchmark DNS predictions for the MDHB and other district health boards. We encourage conducting additional qualitative research to investigate the root cause of DNS issues and potential solutions. Addressing DNS using better strategies potentially can result in better utilization of health care resources and improve health equity.

(JMIR Med Inform 2024;12:e48273) doi:10.2196/48273

KEYWORDS

Did Not Show; Did Not Attend; machine learning; prediction; decision support system; health care operation; data analytics; patients no-show; predictive modeling; appointment nonadherence; health equity

Introduction

Adding to the existing pressures on the health care system [1,2], further substantial disruptions are caused when patients fail to attend their prescheduled appointments [3]. This is defined as Did Not Show (DNS), which is a scheduled but not utilized clinical appointment that patients failed to attend without canceling or rescheduling. This phenomenon is also known as Did Not Attend (DNA) or Failed To Attend (FTA). Causes include the patient forgetting about their appointment, miscommunication [4], logistical difficulties, appointment scheduling conflicts, and family/work commitments [3,5].

DNS can adversely affect patients' well-being, cause them and the system financial stress, and disturb health care operations and systems. Globally, DNS has an overall rate of 23%, with a wide geographical variation (13.2% in Oceania, 19.3% in

Europe, 23.5% in North America, 27.8% in Asia, and 43% in South America [6]). DNS is expensive for health systems; for example, estimated annual losses amounting to £790 million (over US $1 billion) were found in the United Kingdom [7] and $564 million in the United States [8]. It affects both primary and secondary health care [9], although secondary care losses are higher.

Patients mostly fail to comply with their clinical appointments when symptoms become less severe or unnoticeable [10,11], which might deteriorate underlying syndromes [12,13]. Patients are more likely to demand immediate medical attention when contracting serious health issues or require acute and emergency care if they miss scheduled health care appointments [12,14-16].

Eliminating DNS is hard to achieve, and its adverse effects necessitate methods and approaches for managing DNS such as sending digital reminders by text, phone, and email [17,18]. These approaches have not been very effective, as they are time-consuming and costly, and the health care system still faces DNS issues. Overbooking [3,19], open access [20], and DNS penalty approaches have also been used to enhance clinical slot utilization but can cause longer waiting times for patients and overtime for clinical staff [21].

Inspired by the success of artificial intelligence (AI) in different sectors, including health care [22,23], we considered the application of AI for DNS management via predicting the probability of DNS appointments [13,19,24,25]. AI and its subset techniques, such as machine learning (ML), are powerful for extracting cognitive insights from massive amounts of data [26,27].

The predicted DNS probabilities proved to be successful in providing the required information for DNS management [25] and supporting health care managers in making informed decisions for prioritizing patients and delivering clinical assistance. This enables health care providers to reschedule and reuse limited clinical resources for urgent cases while also expanding access to health care services for patients from diverse backgrounds, thereby promoting health care equity.

Therefore, clinical capabilities and medical resources can be used more effectively and efficiently, decreasing patients' wait times, increasing their satisfaction, and enhancing health productivity.

Most studies concerned with predicting DNS have mainly comprised small data sets or specific groups of people to develop models for DNS learning and prediction; however, DNS tends to be varied across populations. For example, longer distances to a medical facility increase DNS [8], but this finding was contradicted in another study [28]. Likewise, patients with chronic illnesses adhere to their scheduled appointments [13], while other studies [29] have shown that patients with more severe diseases have a higher DNS rate. Even within a single medical organization, DNS factors vary across different clinics [14]. These examples highlight the inconsistent nature of DNS predictors, showcasing the complexity of predicting tasks in this domain. Such variations pose challenges in creating a universal formula or model to effectively address DNS prediction issues on a global scale.

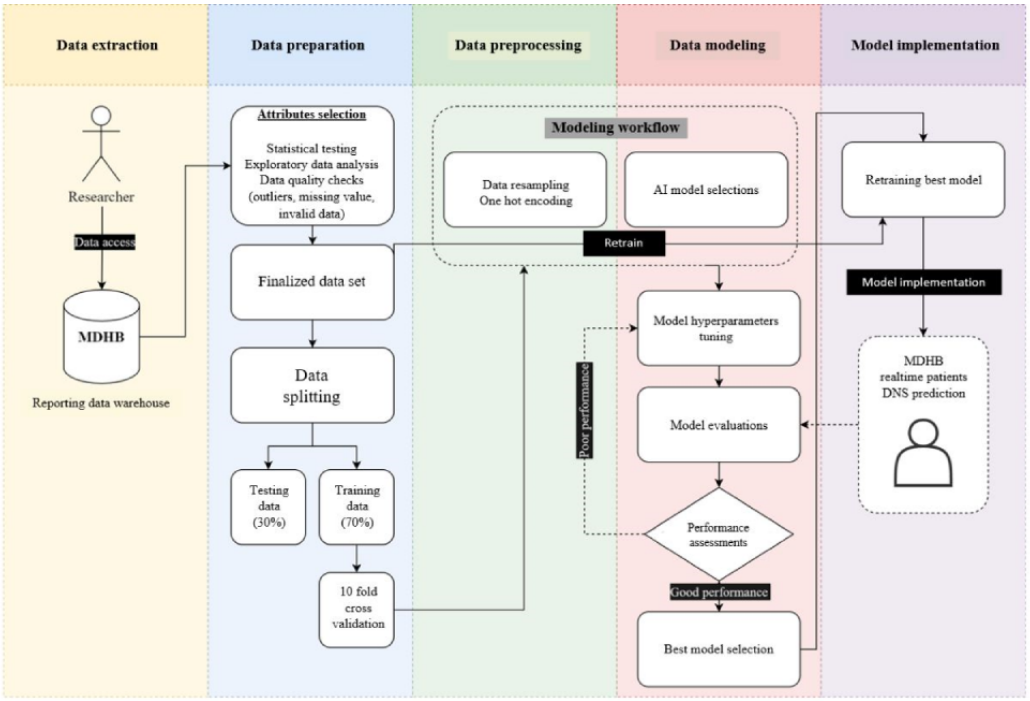
Considering the very limited DNS research in New Zealand and the complexity of developing a general DNS predictive model, we concentrated on the DNS issue in the MidCentral District Health Board (MDHB) hospital as a proof of concept. MDHB is located in the center of the North Island, New Zealand, covering a land area of over 8912 km^2 and with a population of over 191,100 people. In this region, about 18% of people are aged 65 years or older, with over 20% being Māori, and a higher proportion than the national average resides in more deprived areas [30]. These demographic factors could lead to inequity in access to health care services. To support MDHB in addressing health equity and providing additional support for patients, this study aimed to develop ML models and compare their performance in predicting the probabilities of future DNS appointments at MDHB. This study utilized a data set spanning 5 years of collected data.

Methods

Overview

Our research was organized into the following phases (Figure 1)**.** The initial phase involved *data extraction*, defining the data set to be used, and outlining the data extraction process. The *data preparation* phase involved conducting exploratory data analysis (EDA) to profile data and exclude irrelevant observations from the research. Subsequently, the data set was split into 2 parts—70% (454,831 records) for training and 30% (194,927 records) for testing. To avoid data linkage, the training and testing data sets were not mixed during the ML modeling phase. Moreover, the training set underwent a 10-fold cross-validation strategy to prevent bias as much as possible and fully utilize its limited training information. Next, the *data preprocessing* phase involved cleaning and transforming the cross-validation sets, ensuring that the training set was ready for the data modeling stage. A 10-fold cross-validation resampling strategy was applied to further optimize the utilization of the 70% training data. In the *data modeling* phase, we used 3 ML algorithms and tuned their hyperparameters to identify the best performance among the algorithms. Finally, in the *model evaluation* phase, various evaluation metrics were employed to determine the best-performing ML model for DNS prediction.

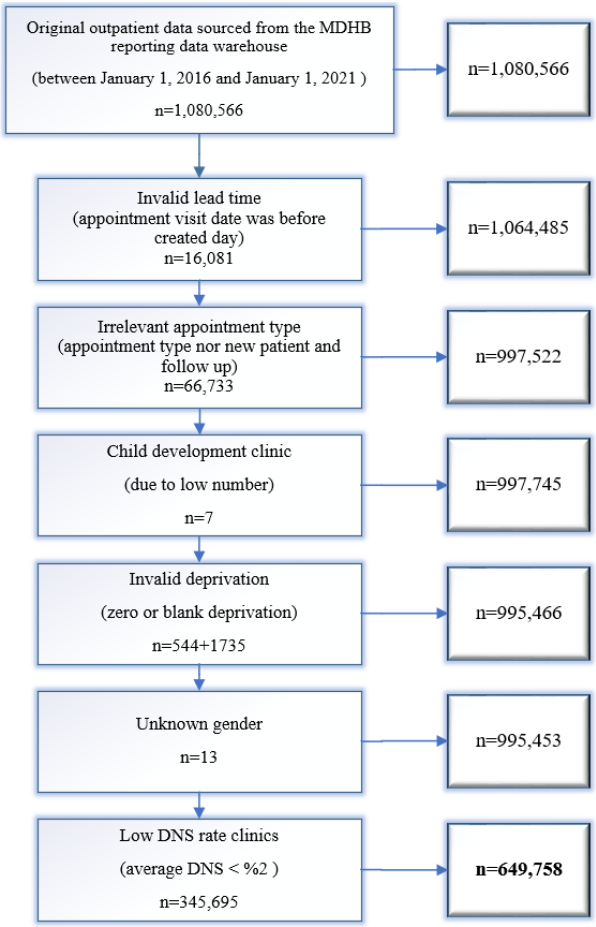
Figure 1. Research flow and procedure. AI: artificial intelligence; MDHB: MidCentral District Health Board.



Data Access and Extraction

Our data were sourced from MDHB reporting SQL farm and contained only outpatient visits with no link to other data sets. This significantly mitigated risks related to patient reidentification. Data deidentification and encryption were applied before data access, and New Zealand National Health Index numbers were encrypted to protect patients' privacy. We acquired 1,080,566 outpatient visit records from 38 clinics between January 1, 2016, and December 31, 2020, satisfying the research requirements with almost 57,000 DNS incidents (5% of the entire data set). The steps of data exclusion are presented in Figure 2. Because not many missing records were identified in the data sets, those with missed values were directly excluded.

Figure 2. Research data exclusion. DNS: Did Not Show; MDHB: MidCentral District Health Board.



Ethical Considerations

This study received ethics approval from the Auckland University of Technology (AUT; 20/303) and MDHB (2020.008.003), following which data access to the MDHB reporting data warehouse was granted.

Data Preparation

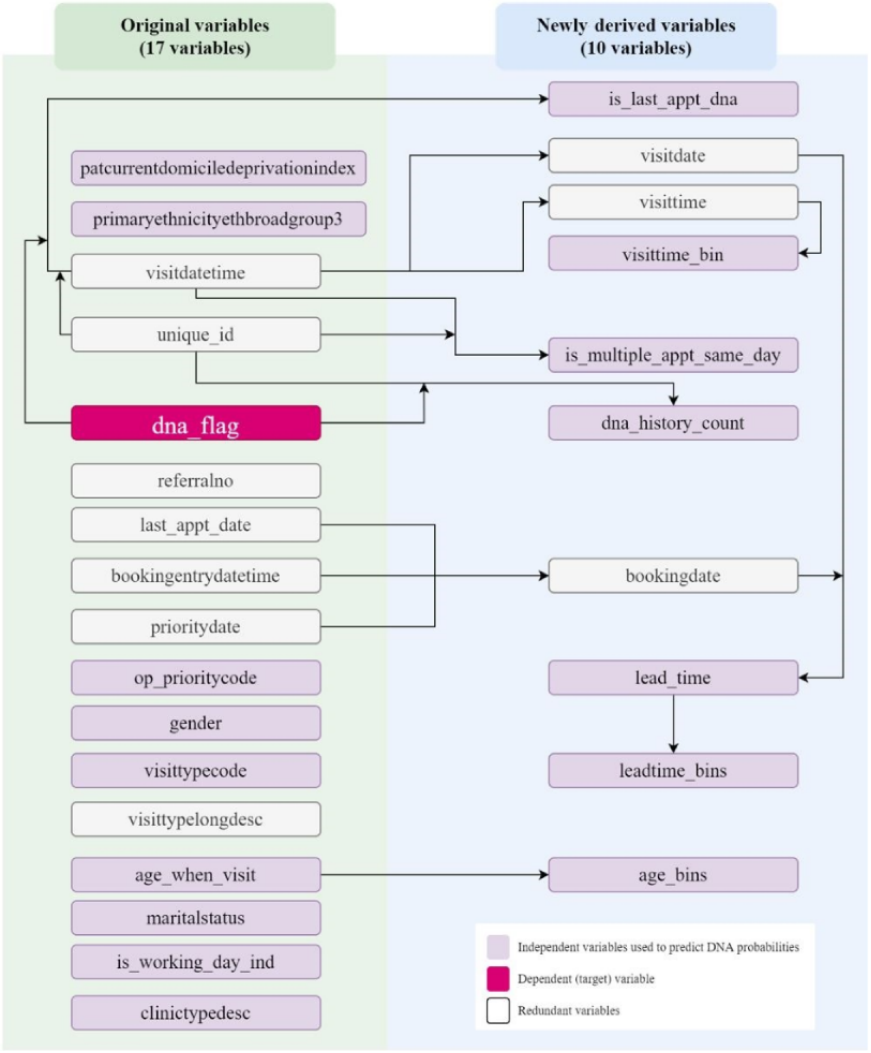
Phase Description

In this phase, understanding the data was important to adequately prepare them for the experiments. The data preparation process included data transformation and derivation (Figure 3). Following suggestions from the literature, new research variables were derived and introduced because some valuable DNS predictors were absent in the MDHB data set. For example, no direct information was available on the patients' DNS history [21,31], appointment lead time [31,32], or latest appointment DNS outcome [13]. The lead time was calculated by comparing the difference in days between the appointment creation date and the visit date. Appointments with longer lead times were expected to have greater DNA probability than those with shorter lead times [29].

Therefore, to better understand patient behavior and DNS patterns, we derived 10 new variables on top of the original variables (Figure 3). These attributes were introduced to support us in understanding when patients were more likely to miss their appointments in general and to identify regular nonadherent patients.

Initially, we extracted a data set with 17 columns and over 1 million records (Multimedia Appendix 1). Informed by the literature review [14,29,31-33], we derived and introduced another 10 variables on top of the original data and increased the data columns to 27. Among all the variables, 16 (59%) were used for ML modeling, and the redundant ones were excluded. The *dna_flag* attribute was the dependent (target) variable. Figure 3 demonstrates the original variables in addition to 10 newly derived ones.

Figure 3. Variable transformation and derivation visualization.



Cardinality Reduction

We conducted a cardinality reduction analysis to reduce variable categories with low frequency and small samples. The data set mostly included categorical variables, with numeric variables being rare. Each categorical level is called a cardinality, which means how many distinct values are in a column. In our data set, some categorical variables had fewer levels, such as patient gender—M (male), F (female), and U (unknown)—while others had hundreds of variations, such as suburbs or diagnosis codes.

Developing ML models often involves numerous categorical attributes, necessitating examination of the variables' cardinality, as most ML algorithms are distance-based and require converting categorical variables to numeric values. Categorical variables with high cardinality levels will derive massive new columns and expand the data set. This expansion increases model complexity, elevates computational costs, and decreases model generalization, which makes handling the data set challenging [34]. Therefore, we investigated the cardinality of our research variables and deployed a reduction strategy accordingly.

Cardinality reduction analysis was conducted to reduce the number of categories within variables with low frequency and small sample sizes. Following suggestions from the literature, new research variables were also derived and introduced, including patients' prior DNS history [14,16,21] and the appointment lead time [14,16,29,32].

Statistical Test

The chi-square test was used for analyzing homogeneity among different groups within variables [35] and for testing the independence between categorical variables [36]. The chi-square statistics (χ^2) and their *P* values were calculated to investigate whether different levels of a variable contributed differently to DNS events.

The confidence level (α=.05) was adopted as the *P* value threshold in the chi-square test. A *P* value less than .05 provided enough confidence to reject the null (H_0) hypothesis and accept the alternative hypothesis (H_A). The tested categorical variable was associated with DNS events [36]. Hence, we may consider using it for future prediction.

After the data preparation process, 16 variables were selected to predict the target *dna_flag*. Among them, 12 modeling predictors were nominal variables, including binary variables (Multimedia Appendix 2). We, therefore, conducted the chi-square (χ^2) statistical test to investigate the relationship

between those predictors and DNS events (Table 1). The chi-square was calculated using the following equation, where O and E are observed and expected values [36,37]:

After preparing the data set and before developing the ML models, an EDA was conducted to gain a deeper understanding of the research data landscape. EDA is a fundamental data analysis required before hypothesis and modeling formulation [38]. Its findings can be used to verify misleading models at a later stage [38] and reveal unexpected patterns [39]. The EDA helped uncover patients' DNS patterns through data aggregation and data visualization analysis. Finally, the EDA findings were validated against the ML model outcomes to verify their accuracy.

Table 1. Chi-square test on categorical variables.

Categorical variables	Chi-square statistic	Chi-square *P* value
dna_history_count	118,461	<.01
is_last_appt_dna	77,600	<.01
Clinictypedesc	35,201	<.01
age_bins	34,810	<.01
primaryethnicityethbroadgroup3	17,098	<.01
leadtime_bins	11,048	<.01
maritalstatus_group	10,527	<.01
visit_type_group	3525	<.01
visittime_bin	3447	<.01
patcurrentdomiciledeprivationindex	2655	<.01
is_multiple_appt_same_day	1913	<.01
op_prioritycode_group	1,496	<.01
is_working_day_ind	1,244	<.01
Gender	4	.06

Data Preprocessing

Due to the high number of categorical variables in our data set, the one-hot encoding technique was used in the preprocessing phase. Because distance-based algorithms can only deal with numerical values, in the cardinality reduction section, we used the one-hot encoding method to convert our categorical variables to numbers. After the conversion, different variables were introduced to our training data set, also known as indicator variables. For example, the variable gender derived 3 variables, *gender_male*, *gender_female*, and *gender_unknown*. Each of those variables can have a value of either 1 or 0.

As the predictive performance of classifiers is highly impacted by the selection of the hyperparameters [40], we conducted hyperparameter tuning to optimize our algorithms' learning process. We further optimized this process using the Grid Search method to boost the performance of our chosen models. Table 2 outlines specific details regarding the hyperparameters utilized.

Table 2. Hyperparameter tuning of the data modeling.

Models and hyperparameters	R package	Range	Purpose
Logistic regression			
penalty	Glmnet	1e-10- 1	Total amount of regularization used to prevent overfit and underfit
Random forest			
Trees	Ranger	300- 1000	Number of trees in the forest
Min_n	Ranger	3-10	Minimum amount of data to further split a node
Mtry	Ranger	3-5	Maximum number of features that will be randomly sampled to split a node
XGBoost[a]			
Trees	XGBoost	300-1000	Number of trees in the forest
Min_n	XGBoost	3-10	Minimum amount of data to further split a node
mtry	XGBoost	3-10	Maximum number of features that will be randomly sampled to split a node
tree_depth	XGBoost	3-12	Maximum depth of the tree

[a]XGBoost: Extreme Gradient Boosting.

Data Modeling

Addressing the imbalanced data set posed the main data modeling challenge. The annual DNS rate for MDHB was around 5%, which means 95% of the appointments were attended visits. This imbalance significantly affected the accuracy of our ML model in predicting attended cases. To tackle this issue, various internal and external strategies exist [41,42]. In this study, we employed an external approach that involved utilizing standard algorithms intended for a balanced data set but applying resampling techniques to the trained data set to reduce the negative impact caused by the unequal class. Our focus was on the resampling strategy, known for its effectiveness in handling imbalanced classification issues and its portability [42].

The resampling strategy involved 2 methods: (1) oversampling, where the size of the minority class is increased randomly to approach the majority class in a class-imbalance data set [43,44]; and (2) undersampling, where the size of the majority class decreases randomly to align with the minority class [43,44]. This strategy falls under both the oversampling and undersampling categories. Given the lack of definitive guidance on the effectiveness of these methods [42-44], we adopted both and compared their results.

Since we dealt with a binary classification prediction problem, supervised and classification algorithms were selected. Algorithms with good interpretability were also considered to explain which predictive variables influence DNS prediction more significantly. In a study concerning variable importance, tree-based models, such as random forest (RF) and gradient-boosted decision trees, were shown to inherently possess features that measure variable importance [45].

For the imbalanced data set, we used ensembling methods due to their proven advantages [46,47]. The following algorithms were chosen for developing DNS prediction models: logistic regression (LR), RF, and Extreme Gradient Boosting (XGBoost).

LR was chosen because it is a suitable analysis method across multiple fields for managing binary classification [48]. Our research concerned a supervised classification problem to predict whether a future outpatient appointment will become a DNS visit. With the response variable (*dna_flag*) offering dichotomous outcomes—either yes (1) or no (0)—LR stood as a fitting choice due to its proficiency in predicting binary outcomes and its established effectiveness in prior studies [7,13,33,49]. Tree-based ensembling algorithms were also chosen for their proven ability to deal with imbalanced data sets and model explainability [46,47]. RF can effectively handle combining random resampling strategies in imbalanced prediction. Tree-ensembling methods have more advanced prediction ability than a single model because they integrate prediction strength from several base learners [50].

Model Implementation and Evaluation

We used 10-fold cross-validation for model selection and bias reduction. The hyperparameters were tuned to boost each classifier's performance. We followed suggestions from the literature suggestions to use sensitivity, specificity, and the area under the receiver operating characteristic (AUROC) curve to quantify the models' prediction strength for the imbalance problem prediction.

During this phase, we used the testing data to validate the best predictive model chosen based on the model evaluation criteria. For this study, data before 2021 were used in the data modeling process. We coordinated with MDHB to access outpatient appointments from 2021 for model validation. Specifically, we used both weekly and monthly data for prediction, comparing these with actual appointment outcomes to validate the model. The benefit of using a new data set for validation was to assess model bias and goodness of fit outside the research environment. Positive performance and high prediction accuracy would

indicate potential real-life implementation of our research model after further investigation.

Results

Our study only included new patients and follow-up appointments. Therefore, we analyzed DNS costs limited to new patient and follow-up outpatient services over the last 5 years. The MDHB provided us with costing information for 34 different departments, and we calculated the DNS cost for each department (Table 3). In 2020, there were 2812 new patient DNS visits and 6240 follow-up DNA visits causing a loss of at least $2.9 million (US $1.8 million) at MDHB. More information regarding this calculation is provided in Multimedia Appendix 3 [51].

Each department was assigned a corresponding outpatient appointment price for a new patient and follow-up outpatient appointment services. We aggregated the total DNS occurrences of new patients and follow-up appointments, multiplying corresponding unit prices to quantify their financial impact. For instance, in 2020, there were 301 new patients and 745 follow-up patients who missed their scheduled bookings, which caused a revenue loss of $300,442 (US $190,000) in the orthopedics department.

Although the initial research expected to address the DNS issue for all outpatient clinics and patients at the MDHB, due to the broad scope of the DNS, we concentrated on clinics with a higher percentage of DNS and narrowed down the research scope to prioritize workloads. To successfully build a model for our focused patient groups, we eliminated as many irrelevant data points as possible. Then, data used for the model training were more fit for purpose for the high-needs population.

The modeling data set was created using 649,758 records and 17 columns (Figures 1 and 3). We developed ML models based on LR, RF, and XGBoost algorithms, with hundreds of hyperparameter combinations in our data modeling. To evaluate the models' prediction performance, accuracy, sensitivity, specificity, AUROC curves, and cost (computation time) were calculated (Table 4). The aim was to identify the best model and hyperparameters that resulted in optimal sensitivity and AUROC performances. Model prediction accuracy is critical; however, it was not a primary concern in this research as we dealt with an imbalanced data set [52].

Table 4 presents a summary comparison of the models' performance. As shown in the table, the LR-based model was the fastest and RF the slowest in terms of computation time. LR had the lowest AUROC scores (ie, the low DNS events prediction accuracy), while RF and XGBoost had a similar area under the curve (AUC) performance (around 0.92).

The undersampling strategy significantly improved our models' sensitivity. Sensitivity was chosen over accuracy because we were dealing with an imbalanced data set [52]. Sensitivity quantified the models' ability to correctly predict positive (DNS) cases that help detect high-risk DNS patients. RF and XGBoost had a very close sensitivity of 0.82. However, considering the computation cost factor, XGBoost had the lowest modeling time. XGBoost with undersampling was our best ML model for the DNS prediction. Its ROC curve is illustrated in Figure 4.

A further investigation was also performed to identify the top predicting factors for each model (Multimedia Appendix 4). The purpose of calculating variable significance scores was not to plug them into a calculation formula but to showcase which variables were more relatively critical in calculating the risk of DNS. Variable importance is critical to AI model development, as variables do not contribute evenly to the final prediction. Therefore, we focused on the most influential predictors and excluded irrelevant ones by scoring the variables' prediction contributions [53]. Variable importance is a measurement quantifying the relationship between an independent variable and the dependent [46].

The results shown in Multimedia Appendix 4 matched the chi-square statistical test results (Table 1). The leading factors were determined and selected using the variable (feature) importance. It was evident that the *dna_history_count* variable was the most influential predictor following *is_last_appt_dna*, *age_when_visit*, and *lead_time*. Additionally, *ethnicity* played an important role in constructing the XGBoost model for the DNS prediction.

We also aggregated outpatient appointment data and ranked the observed DNS rate of all outpatient clinics (Multimedia Appendix 5). We carried out this analysis to initiate an understanding of how disease type might influence the DNS rate.

Table 3. DNS[a] costs in 2020 at the MDHB[b] hospital[c].

Clinics	NP[d] DNS count	NP DNS price	Total NP DNS cost	FU[e] DNS count	FU DNS price	Total FU cost	Total DNS cost
Orthopedics	301	$346	$104,143	745	$263	$196,299	$300,442
Diabetes	90	$452	$40,658	576	$307	$176,643	$217,302
Ophthalmology	221	$239	$52,776	874	$174	$152,322	$205,099
Pediatric medicine	124	$600	$74,366	327	$395	$129,271	$203,637
Ear nose throat	253	$358	$90,571	367	$269	$98,744	$189,316
Gynecology	177	$403	$71,322	386	$280	$108,124	$179,446
Hematology	75	$632	$47,389	232	$348	$80,834	$128,223
Cardiology	109	$490	$53,397	245	$299	$73,259	$126,656
Radiation oncology	42	$505	$21,194	350	$293	$102,652	$123,846
General surgery	147	$387	$56,856	208	$309	$64,369	$121,225
Audiology	268	$214	$57,302	272	$214	$58,157	$115,459
Neurology	153	$617	$94,408	38	$400	$15,204	$109,612
Gastroenterology	68	$506	$34,393	186	$362	$67,401	$101,794
Medical oncology	18	$650	$11,703	229	$360	$82,327	$94,030
Dental	136	$244	$33,132	193	$244	$47,019	$80,151
Renal medicine	5	$559	$2,793	181	$344	$62,201	$64,995
Respiratory lab	38	$479	$18,192	121	$347	$42,021	$60,213
Obstetrics	101	$227	$22,906	143	$227	$32,431	$55,337
Respiratory sleep	20	$271	$5412	153	$271	$41,403	$46,815
Urology	65	$357	$23,178	85	$274	$23,253	$46,432
Dietetics	93	$175	$16,302	168	$175	$29,449	$45,751
General medicine	44	$517	$22,747	69	$322	$22,200	$44,948
Respiratory	39	$479	$18,671	70	$347	$24,309	$42,980
Dermatology	66	$316	$20,877	60	$236	$14,174	$35,051
Oral and maxillofacial	23	$296	$6799	124	$203	$25,185	$31,984
Endocrinology	25	$525	$13,127	34	$332	$11,284	$24,411
Rheumatology	18	$647	$11,643	31	$345	$10,693	$22,336
Plastic surgery (excluding burns)	18	$296	$5321	69	$203	$14,014	$19,335
GI[f] endoscopy	0	$506	$0	52	$362	$18,843	$18,843
Community pediatrics	20	$600	$11,994	10	$395	$3953	$15,948
Infectious diseases	7	$738	$5169	19	$534	$10,152	$15,321
Neurosurgery	1	$507	$507	29	$448	$12,990	$13,496
Podiatry	17	$207	$3522	47	$207	$9737	$13,259
Aged ATR[g] health	18	$244	$4394	35	$244	$8545	$12,939
Under 65 ATR	3	$244	$732	5	$244	$1221	$1953
Cardiothoracic	0	$573	$0	4	$425	$1698	$1698
Anesthetics	9	0	$0	3	$0	$0	$0

[a]DNS: Did Not Show.

[b]MDHB: MidCentral District Health Board.

[c]A currency exchange rate of NZD $1=US $0.61 is applicable for the listed costs.

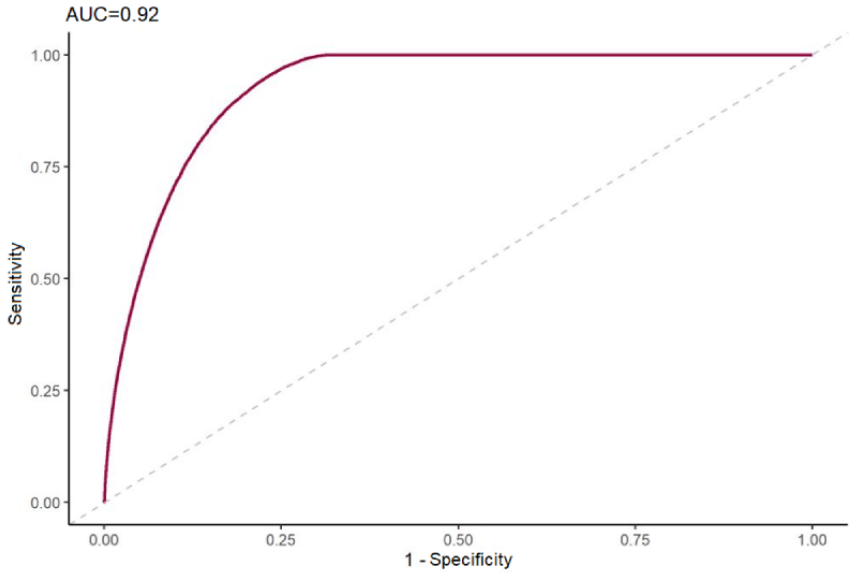
[d]NP: new patient.
[e]FU: follow-up.
[f]GI: gastrointestinal.
[g]ATR: assessment, treatment, and rehabilitation.

Table 4. Comparison of the ML[a] models' performance.

Classifier and resampling strategy	Sensitivity	Specificity	AUC[b]	Accuracy	Modeling cost
Logistic regression					
Undersampling (under_ratio=2)	0.5146	0.9227	0.8474	0.8897	Less than 1 hour (5 minutes)
Oversampling (over_ratio=0.5)	0.5091	0.9247	0.8592	0.8911	Less than 1 hour (14 minutes)
Random forest					
Undersampling (under_ratio=2)	0.8243	0.8524	0.9236	0.8501	Over 8 hours (8.4)
Oversampling (over_ratio=0.5)	0.5940	0.9260	0.9220	0.8990	Over 137 hours
XGBoost[c]					
Undersampling (under_ratio=2)	0.8278	0.8490	0.9239	0.9117	Over 4 hours (4.8)
Oversampling (over_ratio=0.5)	0.8297	0.8549	0.9267	0.8529	Over 51 hours (51.83)

[a]ML: machine learning.
[b]AUC: area under the curve.
[c]XGBoost: Extreme Gradient Boosting.

Figure 4. The receiver operating characteristic (ROC) of the best classifier, Extreme Gradient Boosting (XGBoost). AUC: area under curve.



Discussion

Principal Findings

Our results are comparable to similar previously published analyses [9], although the AUC for XGBoost was slightly higher in our case. This may be due to the data selection and local characteristics. We initially built a generic DNS prediction model for all outpatient clinics at MDHB. However, in light of the literature and DNS complexity, the project scope was narrowed down to clinics with higher DNS rates. As discussed previously in this paper, we excluded irrelevant and missed data, invalid lead time appointments, and clinics with very low DNS rates. This approach improved the ML models' performance and made sense from an operational perspective. The developed models provided insights useful for understanding the contributing factors for DNS. We found that patient DNS history, appointment characteristics, work commitments, and socioeconomic status substantially contributed to DNS events.

Patient DNS History

Understanding patients' DNS history was crucial for predicting future DNS patterns (Table 5) and developing the ML models. This also aligned with the chi-square test results (Table 1), which ranked the *dna_history_count* and *is_last_appt_dna* variables as the most important factors. Total DNS counts and the latest appointment's DNS outcome are pivotal for calculating

the probabilities of future DNS occurrences. These factors are consistent with the findings in the literature [14-16,21,32,54].

Managing DNS involves identifying patients with low adherence to scheduled visits for additional attention. Centralizing and managing DNS history can provide a comprehensive view, preventing data silos or gaps. Centralized monitoring can enhance the visibility of recurring DNS incidents and proactively alert clinicians of potential DNS cases. Our models account for changes in DNS behavior. To reduce the prediction bias, we screen for the most recent appointment DNS outcome (*is_last_appt_dna*).

Table 5. Top prediction variables in the developed ML[a] models.

Algorithm and variable importance ranking	Undersampling model	Oversampling model
Logistic regression		
1	dna_history_count	dna_history_count
2	is_working_day	is_working_day
3	is_last_appt_dna	is_multiple_appt_same_day
4	is_multiple_appt_same_day	is_last_appt_dna
5	lead_time	lead_time
Random forest		
1	dna_history_count	dna_history_count
2	is_last_appt_dna	age_when_visit
3	lead_time	lead_time
4	age_bins	is_last_appt_dna
5	clinic_type_desc	clinic_type_desc
XGBoost[a]		
1	dna_history_count	dna_history_count
2	is_last_appt_dna	age_when_visit
3	age_when_visit	is_last_appt_dna
4	lead_time	ethnicity
5	Ethnicity	lead_time

[a]XGBoost: Extreme Gradient Boosting.

Appointment Characteristics

Certain appointments expected more nonadherence, with distinct predictors related to appointment characteristics such as "working day" and "high lead time." Longer lead times correlated with increased DNS probability, while appointments on working days were more prone to DNS than nonworking days. These findings align with reports from [33,54,55] and emphasize the significant impact of appointment lead time on DNS prediction, as also indicated in [8,14,16,32,33,54]. This underscores how appointment characteristics directly affect DNS outcomes immediately after scheduling. Therefore, incorporating ML-predicted DNS risk estimations during appointment scheduling could automatically flag higher DNS probability for proactive management.

Furthermore, our analysis of the *op_priorityc*ode variable (Multimedia Appendix 1) indicated that, in general, patients with more serious health conditions were more likely to attend their appointments. This observation is reflected in Multimedia Appendix 5, which compares the DNS rates of different clinics with the overall average DNS rate of 0.053% (depicted red line). For example, patients visiting the audiology clinic had a potential DNS rate of 19.1% compared to a 0.9% DNS rate for the radiation oncology clinic. Our analysis of the *op_priorityc*ode variable was based on categorical data types reflecting appointment urgency and not based on a detailed analysis of each patient's diagnosis.

Work Commitments

Our findings suggest that patients struggled to adhere to appointments on working days or during working hours. Younger adults, particularly those between 20 and 30 years of age, had higher DNS rates due to work commitments, while older adults aged 65 years and above rarely missed their visits.

Furthermore, the XGBoost-based model highlighted that being single was an indicator of DNS visits (Figure 4). This could relate to time constraints among young professionals, a finding consistent with other studies [8,28,33,56]. For this group, a targeted reminder system could be developed to concentrate on appointments with higher DNA probability compared to the

DNS risk threshold. Consequently, the population-based reminding system could help optimize resource allocation, including staff efforts and costs.

Socioeconomic Status

We explored the deprivation index and clustered patient populations by using their ethnicity (Multimedia Appendix 6). Our findings indicated a strong association between European and Māori ethnicities and DNS outcomes, ranked among the top 5 predicting factors (Multimedia Appendix 4). Māori and Pacific populations had the highest DNS rates, in line with other research findings [56], while the European ethnicity had the lowest DNS rates. Māori and Pacific populations tended to reside in areas characterized by higher deprivation rates, whereas the percentage of other ethnicities living in higher deprivation regions decreased when the deprivation index increased.

In New Zealand, Māori and Pacific ethnical groups required increased health care attention [57] to ensure equity in the health care system. As indicated in Table 6, a larger proportion of these ethnic groups are situated in suburbs and areas with higher deprivation indexes (such as 8, 9, and 10) [58]. The higher deprivation index was also a strong indicator of socioeconomic deprivation geographically [58]. According to the New Zealand Index of Deprivation, neighborhoods with higher deprivation were more likely to experience adverse living conditions such as damp, cold, and crowded housing.

Moreover, regions with higher deprivation exhibit higher rates of unemployment, increased dependence on benefits, and more single-parent families [58]. Consequently, these living conditions and income disparities made patients living in these regions more susceptible to illness, while also encountering more barriers and obstacles in addressing their medical needs.

At MDHB, dedicated working groups were established to support Māori and Pacific patients in attending their scheduled hospital appointments. Our research reiterates the importance and necessity of those working groups, acknowledging the value of their work. Moreover, our model can support them further by providing tangible DNS probability scores to prioritize patients who require additional attention and support.

Table 6. Percentage of population residing at each deprivation level [58].

Deprivation level	Māori, n (%)	Pacific, n (%)	European, n (%)	Asian, n (%)	Other, n (%)
1	3113 (7)	293 (1)	37,314 (86)	2077 (5)	835 (2)
2	4951 (9)	429 (1)	46,405 (85)	1470 (3)	1071 (2)
3	6367 (13)	489 (1)	42,565 (84)	613 (1)	821 (2)
4	14,736 (14)	1747 (2)	84,728 (79)	4574 (4)	1593 (1)
5	14,400 (13)	3398 (3)	83,568 (77)	6015 (6)	1590 (1)
6	14,103 (15)	1759 (2)	74,351 (79)	2974 (3)	1248 (1)
7	13,442 (17)	3601 (5)	58,187 (75)	1858 (2)	870 (1)
8	36,843 (19)	5402 (3)	148,605 (75)	5434 (3)	1988 (1)
9	40,642 (24)	7324 (4)	111,319 (67)	5443 (3)	2442 (1)
10	31,998 (35)	6283 (7)	52,064 (56)	1610 (2)	521 (1)

Operational and Managerial Implications

The total DNS loss incurred by the MDHB hospital was around $2.9 million (US $1.8 million) in 2020. Notably, we observed that clinics with less life-threatening diseases (diabetes, audiology, and dental) had higher DNS rates. Considering our use of MDHB data, we expect to identify similar patterns in other district health boards for which the same DNS predicting factors can be applied for DNS management.

While the primary objective of our research was to calculate DNS risk for promoting health equity, we believe that leveraging DNS prediction can aid in managing limited health care resources more efficiently. By quantifying the DNS probability for future appointments on a scale from 0.00 to 1, clinicians or hospital operation managers can develop more personalized health care services for their patients. This leads to enhancing equity in accessing health care services for a wider population.

The predictions derived can support MDHB managers in designing, planning, and implementing more informed DNS management strategies. For example, a DNS appointments threshold (eg, 0.7) can be set, and all appointments with predicted odds greater than 0.7 can be selected, releasing 70% of resources and allocating some (or all) to the remaining 30% of patients with a higher DNS risk. Potentially, these released resources can subsidize interventions to support attendance. Without DNS prediction, the hospital cannot decide where to focus on solving the DNS problem and must invest money uniformly for every patient, leading to equality rather than equity in health care service access. Equality is not fit for purpose, especially considering the high attendance rate of 95% over the past 5 years, indicating that most patients attend appointments without additional support. However, for more optimum use of health care resources, other policies and guidance for appointment scheduling should be considered [59].

Potential Interventions to Reduce DNS

DNS Suggests Life Hardships

When patients miss medical appointments, it is a critical indicator suggesting they may be experiencing hardships in their lives [15,54,60]. Considering that a higher DNS rate correlates with a higher deprivation index, we can assume that

people residing in these areas may face greater transportation limitations. Moreover, people with severe mental health or addiction issues may not be able to independently visit their doctors [15]. These vulnerable groups require additional and ongoing appointment assistance. Unfortunately, they have been historically disadvantaged and marginalized by the current health care system [61].

The DNS prediction model we developed can help health care practitioners identify patients at higher risk of DNS. Targeted DNS improvement solutions can be designed based on predicted DNS probability, patient demography, and clinical history. This type of application can leverage the DNS prediction model to help identify and deliver patient-centric medical services to patients requiring additional help. Some examples are discussed in the subsequent sections.

Expanding Integrated Health Care Networks

For patients not facing life-threatening illnesses or requiring long-term health management (such as patients with diabetes), expanding services closer to patients might help meet their needs. MDHB could consider deploying clinicians to outsourced sites to supervise practitioners or attend to patients directly. Moreover, increasing collaborations with primary health care networks, promoting nurse-led services, and contracting private specialists can also be viable options for decreasing DNS rates. Developing a one-stop medical hub with multidisciplinary clinics for patients with lower clinical risk could encourage attendance and reduce DNS visits [19]. This is consistent with the New Zealand Ministry's latest health care system reform strategies, which aim to uplift health care equity [61]. The reform emphasizes the establishment of more locality networks in the community, resonating well with our research findings.

After-Hour Appointment Slots

To support young adults who are occupied by daily work, it might be favorable to increase more after-hour service slots in clinics when possible. If more appointment slots can be organized before or after working hours, working professionals may have more chances to adhere to their clinical appointments. Piloting more weekend clinics can also be a choice to meet younger generations' needs. In consonance with our suggestion, the recent New Zealand health care reform also promoted more affordable after-hours services [61]. Additionally, offering transportation assistance and improved wraparound well-being support for patients with a high-risk score could increase attendance. At-home patient visits could also be offered and delivered to patients facing severe transport limitations.

Limitations

Despite the success of our DNS prediction model, we need to acknowledge that it has some limitations. First, our model was trained on 5-year period data from MDHB. The single data source prevented us from exploring other critical dimensions such as household data or beneficiary data. We believe adding those data points would improve the prediction model and discover more patients' DNS patterns.

Furthermore, we pairwise compared the attribute *dna_flag* with other DNS predictor factors. However, future research should consider investigating and analyzing the association between variables and adding further variables to the conditioning set. This expanded analysis would offer deeper insights into patients' DNS behaviors.

Conclusions

To the best of our knowledge, this study represents one of the first attempts in New Zealand to develop ML prediction models supporting DNS management. We successfully developed and tested ML models to predict probabilities of outpatient appointments' DNS. Our selected model had an AUROC of 0.92 and a sensitivity performance of 0.82.

Acknowledgments

The authors would like to thank the New Zealand MidCentral District Health Board (MDHB) for their support of this study. We appreciate the advice, help, and support from the MDHB data analytics team, Dr Richard Fong, and Mr Rahul Alate. Without their contribution, this study would not have been possible.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of variables and their definitions.
[DOCX File , 16 KB - medinform_v12i1e48273_app1.docx]

Multimedia Appendix 2
Data type of original and newly derived variables.
[DOCX File , 15 KB - medinform_v12i1e48273_app2.docx]

Multimedia Appendix 3
Outpatient appointment prices.
[DOCX File , 19 KB - medinform_v12i1e48273_app3.docx]

Multimedia Appendix 4
Leading predicting factors of the best Extreme Gradient Boosting (XGBoost) model.
[DOCX File , 212 KB - medinform_v12i1e48273_app4.docx]

Multimedia Appendix 5
Did Not Show (DNS) rates of all outpatient clinics of the MidCentral District Health Board (MDHB) hospital.
[DOCX File , 329 KB - medinform_v12i1e48273_app5.docx]

Multimedia Appendix 6
Did Not Show (DNS) rates among different deprivation groups and ethnicities.
[DOCX File , 126 KB - medinform_v12i1e48273_app6.docx]

References

1. Tun SYY, Madanian S, Mirza F. Internet of things (IoT) applications for elderly care: a reflective review. Aging Clin Exp Res 2021 Apr;33(4):855-867. [doi: 10.1007/s40520-020-01545-9] [Medline: 32277435]
2. Madanian S. The use of e-health technology in healthcare environment: The role of RFID technology. 10th International Conference on e-Commerce in Developing Countries: with focus on e-Tourism (ECDC); 15-16 April 2016; Isfahan, Iran: IEEE; 2016 Presented at: 10th International Conference on e-Commerce in Developing Countries (ECDC); April 15-16; Isfahan, Iran p. 1-5 URL: https://ieeexplore.ieee.org/document/7492974 [doi: 10.1109/ECDC.2016.7492974]
3. Alaeddini A, Yang K, Reddy C, Yu S. A probabilistic model for predicting the probability of no-show in hospital appointments. Health Care Manag Sci 2011 Jun;14(2):146-157 [FREE Full text] [doi: 10.1007/s10729-011-9148-9] [Medline: 21286819]
4. Kaplan-Lewis E, Percac-Lima S. No-show to primary care appointments: why patients do not come. J Prim Care Community Health 2013 Oct;4(4):251-255. [doi: 10.1177/2150131913498513] [Medline: 24327664]
5. DeFife JA, Conklin CZ, Smith JM, Poole J. Psychotherapy appointment no-shows: rates and reasons. Psychotherapy (Chic) 2010 Sep;47(3):413-417. [doi: 10.1037/a0021168] [Medline: 22402096]
6. Dantas L, Fleck J, Cyrino Oliveira FL, Hamacher S. No-shows in appointment scheduling - a systematic literature review. Health Policy 2018 Apr;122(4):412-421 [FREE Full text] [doi: 10.1016/j.healthpol.2018.02.002] [Medline: 29482948]
7. Blæhr E, Søgaard R, Kristensen T, Væggemose U. Observational study identifies non-attendance characteristics in two hospital outpatient clinics. Dan Med J 2016 Oct;63(10) [FREE Full text] [Medline: 27697132]
8. Davies ML, Goffman RM, May JH, Monte RJ, Rodriguez KL, Tjader YC, et al. Large-scale no-show patterns and distributions for clinic operational research. Healthcare (Basel) 2016 Mar 16;4(1) [FREE Full text] [doi: 10.3390/healthcare4010015] [Medline: 27417603]
9. Nelson A, Herron D, Rees G, Nachev P. Predicting scheduled hospital attendance with artificial intelligence. NPJ Digit Med 2019 Apr 12;2(1):26 [FREE Full text] [doi: 10.1038/s41746-019-0103-3] [Medline: 31304373]
10. Samuels RC, Ward VL, Melvin P, Macht-Greenberg M, Wenren LM, Yi J, et al. Missed appointments: factors contributing to high no-show rates in an urban pediatrics primary care clinic. Clin Pediatr (Phila) 2015 Sep 12;54(10):976-982. [doi: 10.1177/0009922815570613] [Medline: 25676833]
11. Hayton C, Clark A, Olive S, Browne P, Galey P, Knights E, et al. Barriers to pulmonary rehabilitation: characteristics that predict patient attendance and adherence. Respir Med 2013 Mar;107(3):401-407 [FREE Full text] [doi: 10.1016/j.rmed.2012.11.016] [Medline: 23261311]
12. French LR, Turner KM, Morley H, Goldsworthy L, Sharp DJ, Hamilton-Shield J. Characteristics of children who do not attend their hospital appointments, and GPs' response: a mixed methods study in primary and secondary care. Br J Gen Pract 2017 Jul;67(660):e483-e489 [FREE Full text] [doi: 10.3399/bjgp17X691373] [Medline: 28630057]
13. Goffman RM, Harris SL, May JH, Milicevic AS, Monte RJ, Myaskovsky L, et al. Modeling patient no-show history and predicting future outpatient appointment behavior in the Veterans Health Administration. Mil Med 2017 May;182(5):e1708-e1714. [doi: 10.7205/MILMED-D-16-00345] [Medline: 29087915]
14. Mohammadi I, Wu H, Turkcan A, Toscos T, Doebbeling BN. Data analytics and modeling for appointment no-show in community health centers. J Prim Care Community Health 2018;9:2150132718811692 [FREE Full text] [doi: 10.1177/2150132718811692] [Medline: 30451063]
15. Williamson AE, Ellis DA, Wilson P, McQueenie R, McConnachie A. Understanding repeated non-attendance in health services: a pilot analysis of administrative data and full study protocol for a national retrospective cohort. BMJ Open 2017 Feb 14;7(2):e014120 [FREE Full text] [doi: 10.1136/bmjopen-2016-014120] [Medline: 28196951]
16. Lee G, Wang S, Dipuro F, Hou J, Grover P, Low L. Leveraging on predictive analytics to manage clinic no show and improve accessibility of care. : IEEE; 2017 Presented at: IEEE International Conference on Data Science and Advanced Analytics (DSAA); October 19-21; Tokyo, Japan p. 19-21 URL: https://ieeexplore.ieee.org/document/8259804 [doi: 10.1109/DSAA.2017.25]

17. Orskov ER, Fraser C. The effects of processing of barley-based supplements on rumen pH, rate of digestion of voluntary intake of dried grass in sheep. Br J Nutr 1975 Nov;34(3):493-500. [doi: 10.1017/s0007114575000530] [Medline: 36]
18. Prasad S, Anand R. Use of mobile telephone short message service as a reminder: the effect on patient attendance. Int Dent J 2012 Feb 18;62(1):21-26 [FREE Full text] [doi: 10.1111/j.1875-595X.2011.00081.x] [Medline: 22251033]
19. AlMuhaideb S, Alswailem O, Alsubaie N, Ferwana I, Alnajem A. Prediction of hospital no-show appointments through artificial intelligence algorithms. Ann Saudi Med 2019;39(6):373-381 [FREE Full text] [doi: 10.5144/0256-4947.2019.373] [Medline: 31804138]
20. Kunjan K, Wu H, Toscos TR, Doebbeling BN. Large-scale data mining to optimize patient-centered scheduling at health centers. J Healthc Inform Res 2019 Mar 4;3(1):1-18 [FREE Full text] [doi: 10.1007/s41666-018-0030-0] [Medline: 35415421]
21. Lenzi H, Ben AJ, Stein AT. Development and validation of a patient no-show predictive model at a primary care setting in Southern Brazil. PLoS One 2019 Apr 4;14(4):e0214869 [FREE Full text] [doi: 10.1371/journal.pone.0214869] [Medline: 30947294]
22. Chen T, Madanian S, Airehrour D, Cherrington M. Machine learning methods for hospital readmission prediction: systematic analysis of literature. J Reliable Intell Environ 2022 Jan 30;8(1):49-66. [doi: 10.1007/s40860-021-00165-y]
23. Madanian S, Parry D, Adeleye O, Poellabauer C, Mirza F, Mathew S, et al. Automatic speech emotion recognition using machine learning: digital transformation of mental health. 2022 Presented at: Pacific Asia Conference on Information Systems; July 5-9; Taipei, Taiwan and Sydney, Australia URL: https://aisel.aisnet.org/pacis2022/45/
24. Kurasawa H, Hayashi K, Fujino A, Takasugi K, Haga T, Waki K, et al. Machine-learning-based prediction of a missed scheduled clinical appointment by patients with diabetes. J Diabetes Sci Technol 2016 May;10(3):730-736 [FREE Full text] [doi: 10.1177/1932296815614866] [Medline: 26555782]
25. Barrera Ferro D, Brailsford S, Bravo C, Smith H. Improving healthcare access management by predicting patient no-show behaviour. Decision Support Systems 2020 Nov;138:113398 [FREE Full text] [doi: 10.1016/j.dss.2020.113398]
26. Madanian S, Rasoulipanah HR, Yu J. Stress detection on social network: public mental health surveillance. 2023 Presented at: The 2023 Australasian Computer Science Week; January 31-February 3; Melbourne, Australia p. 170-175 URL: https://dl.acm.org/doi/10.1145/3579375.3579397 [doi: 10.1145/3579375.3579397]
27. Madanian S, Chen T, Adeleye O, Templeton J, Poellabauer C, Parry D, et al. Speech emotion recognition using machine learning — a systematic review. Intell Syst Appl 2023 Nov;20:200266 [FREE Full text] [doi: 10.1016/j.iswa.2023.200266]
28. Hamilton W, Round A, Sharp D. Patient, hospital, and general practitioner characteristics associated with non-attendance: a cohort study. Br J Gen Pract 2002 Apr;52(477):317-319 [FREE Full text] [Medline: 11942451]
29. Eid WE, Shehata SF, Cole DA, Doerman KL. Predictors of nonattendance at an endocrinology outpatient clinic. Endocr Pract 2016 Aug;22(8):983-989. [doi: 10.4158/EP161198.OR] [Medline: 27124692]
30. Living in MidCentral: geographic area and population. Te Whatu Ora Health New Zealand. Wellington, New Zealand URL: https://www.careers.mdhb.health.nz/living-in-midcentral [accessed 2023-10-16]
31. Topuz K, Uner H, Oztekin A, Yildirim M. Predicting pediatric clinic no-shows: a decision analytic framework using elastic net and Bayesian belief network. Ann Oper Res 2017 Apr 4;263(1-2):479-499. [doi: 10.1007/s10479-017-2489-0] [Medline: 44]
32. Lin Q, Betancourt B, Goldstein BA, Steorts RC. Prediction of appointment no-shows using electronic health records. J Appl Stat 2020 Jul;47(7):1220-1234 [FREE Full text] [doi: 10.1080/02664763.2019.1672631] [Medline: 35707022]
33. Fiorillo CE, Hughes AL, I-Chen C, Westgate PM, Gal TJ, Bush ML, et al. Factors associated with patient no-show rates in an academic otolaryngology practice. Laryngoscope 2018 Mar 16;128(3):626-631 [FREE Full text] [doi: 10.1002/lary.26816] [Medline: 28815608]
34. García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. Cham, Switzerland: Springer; 2015.
35. McHugh ML. The chi-square test of independence. Biochem Med (Zagreb) 2013;23(2):143-149 [FREE Full text] [doi: 10.11613/bm.2013.018] [Medline: 23894860]
36. Franke TM, Ho T, Christie CA. The chi-square test. Am J Eval 2011 Nov 08;33(3):448-458. [doi: 10.1177/1098214011426594]
37. The R Project for Statistical Computing.: The R Foundation URL: https://www.r-project.org/ [accessed 2023-10-15]
38. Behrens J, DiCerbo K, Yel N, Levy R. Exploratory data analysis. In: Handbook of Psychology: Research Methods in Psychology. New York, NY: John Wiley & Sons; 2012:2012.
39. Behrens JT. Principles and procedures of exploratory data analysis. Psychol Methods 1997 Jun;2(2):131-160. [doi: 10.1037/1082-989x.2.2.131]
40. Mantovani R, Horváth T, Cerri R, Vanschoren J. Hyper-parameter tuning of a decision tree induction algorithm. : IEEE; 2016 Presented at: 5th Brazilian Conference on Intelligent Systems (BRACIS); October 9-12; Recife, Brazil p. 37-42. [doi: 10.1109/BRACIS.2016.018]
41. Doucette J, Heywood M. GP classification under imbalanced data sets: active sub-sampling and AUC approximation. In: O'Neill M, Vanneschi L, Gustafson S, Alcázar A, Falco I, Cioppa A, et al, editors. Genetic Programming. Berlin, Germany: Springer; 2008:9-23.
42. Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Comput Intell 2004 Jan 28;20(1):18-36 [FREE Full text] [doi: 10.1111/j.0824-7935.2004.t01-1-00228.x]

43. Batuwita R, Palade V. Efficient resampling methods for training support vector machines with imbalanced datasets. : IEEE; 2010 Presented at: 2010 International Joint Conference on Neural Networks (IJCNN); July 18-23; Barcelona, Spain p. 1-8. [doi: 10.1109/IJCNN.2010.5596787]
44. Xu-Ying Liu, Jianxin Wu, Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst, Man, Cybern B 2009 Apr;39(2):539-550. [doi: 10.1109/tsmcb.2008.2007853]
45. Greenwell B, Boehmke B. Variable importance plots—An introduction to the vip package. R J 2020;12(1):343. [doi: 10.32614/rj-2020-013]
46. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 2011 Jul 29;11:51 [FREE Full text] [doi: 10.1186/1472-6947-11-51] [Medline: 21801360]
47. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. University of California Berkeley. 2004. URL: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf [accessed 2023-12-28]
48. Hosmer D, Lemeshow S, Sturdivant R. Applied Logistic Regression. Hoboken, NJ: John Wiley & Sons; 2013.
49. Devasahay SR, Karpagam S, Ma NL. Predicting appointment misses in hospitals using data analytics. Mhealth 2017;3:12 [FREE Full text] [doi: 10.21037/mhealth.2017.03.03] [Medline: 28567409]
50. Sahin EK. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Appl Sci 2020 Jun 30;2(7). [doi: 10.1007/s42452-020-3060-1]
51. Nationwide service framework library online. Te Whatu Ora Health New Zealand. 2023. URL: https://www.tewhatuora.govt.nz/our-health-system/nationwide-service-framework-library/ [accessed 2023-10-14]
52. Nejatian S, Parvin H, Faraji E. Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification. Neurocomputing 2018 Feb;276:55-66. [doi: 10.1016/j.neucom.2017.06.082]
53. Hastie T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer; 2017.
54. Harvey HB, Liu C, Ai J, Jaworsky C, Guerrier CE, Flores E, et al. Predicting no-shows in radiology using regression modeling of data available in the electronic medical record. J Am Coll Radiol 2017 Oct;14(10):1303-1309. [doi: 10.1016/j.jacr.2017.05.007] [Medline: 28673777]
55. Ma N, Khataniar S, Wu D, Ng S. Predictive analytics for outpatient appointments. : IEEE; 2014 Presented at: 2014 International Conference on Information Science & Applications (ICISA); May 6-9; Seoul, South Korea p. 6-9. [doi: 10.1109/ICISA.2014.6847449]
56. Lamba M, Alamri Y, Garg P, Frampton C, Rowbotham D, Gearry R. Predictors of non-attendance at outpatient endoscopy: a five-year multi-centre observational study from New Zealand. N Z Med J 2019 Jun 07;132(1496):31-38. [Medline: 31170131]
57. Maori health. New Zealand Ministry of Health. 2023. URL: https://www.health.govt.nz/our-work/populations/maori-health [accessed 2023-11-05]
58. Atkinson J. Socioeconomic deprivation indexes: NZDep and NZiDe. University of Otago. 2019. URL: https://www.otago.ac.nz/wellington/departments/publichealth/research/hirp/otago020194.html [accessed 2023-11-15]
59. Samorani M, LaGanga L. Outpatient appointment scheduling given individual day-dependent no-show predictions. Eur J Oper Res 2015 Jan;240(1):245-257 [FREE Full text] [doi: 10.1016/j.ejor.2014.06.034]
60. Wilcox A, Levi EE, Garrett JM. Predictors of non-attendance to the postpartum follow-up visit. Matern Child Health J 2016 Nov 25;20(Suppl 1):22-27. [doi: 10.1007/s10995-016-2184-9] [Medline: 27562797]
61. The new health system. New Zealand Department of the Prime Minister and Cabinet. 2021. URL: https://dpmc.govt.nz/our-business-units/transition-unit/response-health-and-disability-system-review/information [accessed 2023-11-15]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
AUROC: area under the receiver operating characteristic
AUT: Auckland University of Technology
DNA: Did Not Attend
DNS: Did Not Show
EDA: exploratory data analysis
FTA: Failed To Attend
LR: logistic regression
MDHB: MidCentral District Health Board
ML: machine learning
RF: random forest
ROC: receiver operating characteristic
XGBoost: Extreme Gradient Boosting

Edited by C Lovis; submitted 17.04.23; peer-reviewed by A Blasiak, D Gartner; comments to author 02.10.23; revised version received 07.11.23; accepted 04.12.23; published 12.01.24.

Please cite as:
Yang Y, Madanian S, Parry D
Enhancing Health Equity by Predicting Missed Appointments in Health Care: Machine Learning Study
JMIR Med Inform 2024;12:e48273
URL: https://medinform.jmir.org/2024/1/e48273
doi:10.2196/48273
PMID:38214974

Original Paper

Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study

Anne de Hond[1,2,3], MSc, PhD; Marieke van Buchem[1,2,3], MSc; Claudio Fanconi[3,4], MSc; Mohana Roy[5], MD; Douglas Blayney[5], MD; Ilse Kant[1,6], MSc, PhD; Ewout Steyerberg[1,2], MSc, PhD; Tina Hernandez-Boussard[3,7,8], MSc, PhD

[1]Clinical AI Implementation and Research Lab, Leiden University Medical Centre, Leiden, Netherlands

[2]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands

[3]Department of Medicine (Biomedical Informatics), Stanford Medicine, Stanford University, Stanford, CA, United States

[4]Department of Electrical Engineering and Information Technology, ETH Zürich, Zürich, Switzerland

[5]Department of Medical Oncology, Stanford Medicine, Stanford University, Stanford, CA, United States

[6]Department of Digital Health, University Medical Centre Utrecht, Utrecht, Netherlands

[7]Department of Biomedical Data Science, Stanford University, Stanford, CA, United States

[8]Department of Epidemiology & Population Health (by courtesy), Stanford University, Stanford, CA, United States

Corresponding Author:
Tina Hernandez-Boussard, MSc, PhD
Department of Medicine (Biomedical Informatics)
Stanford Medicine
Stanford University
1265 Welch Road
Stanford, CA, 94305
United States
Phone: 1 650 725 5507
Email: boussard@stanford.edu

Abstract

Background: Patients with cancer starting systemic treatment programs, such as chemotherapy, often develop depression. A prediction model may assist physicians and health care workers in the early identification of these vulnerable patients.

Objective: This study aimed to develop a prediction model for depression risk within the first month of cancer treatment.

Methods: We included 16,159 patients diagnosed with cancer starting chemo- or radiotherapy treatment between 2008 and 2021. Machine learning models (eg, least absolute shrinkage and selection operator [LASSO] logistic regression) and natural language processing models (Bidirectional Encoder Representations from Transformers [BERT]) were used to develop multimodal prediction models using both electronic health record data and unstructured text (patient emails and clinician notes). Model performance was assessed in an independent test set (n=5387, 33%) using area under the receiver operating characteristic curve (AUROC), calibration curves, and decision curve analysis to assess initial clinical impact use.

Results: Among 16,159 patients, 437 (2.7%) received a depression diagnosis within the first month of treatment. The LASSO logistic regression models based on the structured data (AUROC 0.74, 95% CI 0.71-0.78) and structured data with email classification scores (AUROC 0.74, 95% CI 0.71-0.78) had the best discriminative performance. The BERT models based on clinician notes and structured data with email classification scores had AUROCs around 0.71. The logistic regression model based on email classification scores alone performed poorly (AUROC 0.54, 95% CI 0.52-0.56), and the model based solely on clinician notes had the worst performance (AUROC 0.50, 95% CI 0.49-0.52). Calibration was good for the logistic regression models, whereas the BERT models produced overly extreme risk estimates even after recalibration. There was a small range of decision thresholds for which the best-performing model showed promising clinical effectiveness use. The risks were underestimated for female and Black patients.

Conclusions: The results demonstrated the potential and limitations of machine learning and multimodal models for predicting depression risk in patients with cancer. Future research is needed to further validate these models, refine the outcome label and predictors related to mental health, and address biases across subgroups.

(JMIR Med Inform 2024;12:e51925) doi:10.2196/51925

KEYWORDS

natural language processing; machine learning; artificial intelligence; oncology; depression; clinical decision support; decision support; cancer; patients with cancer; chemotherapy; mental health; prediction model; depression risk; cancer treatment; radiotherapy; diagnosis; validation; cancer care; care

Introduction

Background

Depression in patients with cancer occurs frequently around diagnosis and treatment and has been negatively associated with a patient's prognosis, quality of life, and treatment adherence [1-5]. Despite affecting up to 20% of patients with cancer and far exceeding the prevalence in the general population (8.4% in the United States [6]), depression is underdiagnosed and often untreated [1,3,7-9]. Constrained clinician time and a strong focus on anticancer treatment may contribute to the insufficient identification of patients at risk for depression [10-13]. Early detection of depression in patients with cancer may enable timely mental health support to augment the anticancer treatment.

Clinical decision support tools with artificial intelligence (AI) technologies could synthesize the abundance of data collected during treatment to help clinicians identify which patients may need specific attention and steer additional mental health resources to those at high risk. A recent review [14] of AI models developed for depression risk in primary care [15], elderly care [16,17], and social media posts [18-20] highlights how AI tools have the potential for early identification of mental health issues. However, oncology-specific applications are rare, and those that do exist are developed on selected small samples that may not generalize to clinical care settings [21,22]. This leaves a gap in oncological care for mental health.

Objective

We aimed to develop a prediction model for early identification of patients at risk for depression within the first month of chemo- or radiotherapy treatment. We assessed the relevance of different data modalities for predictive performance in a retrospective cohort study.

Methods

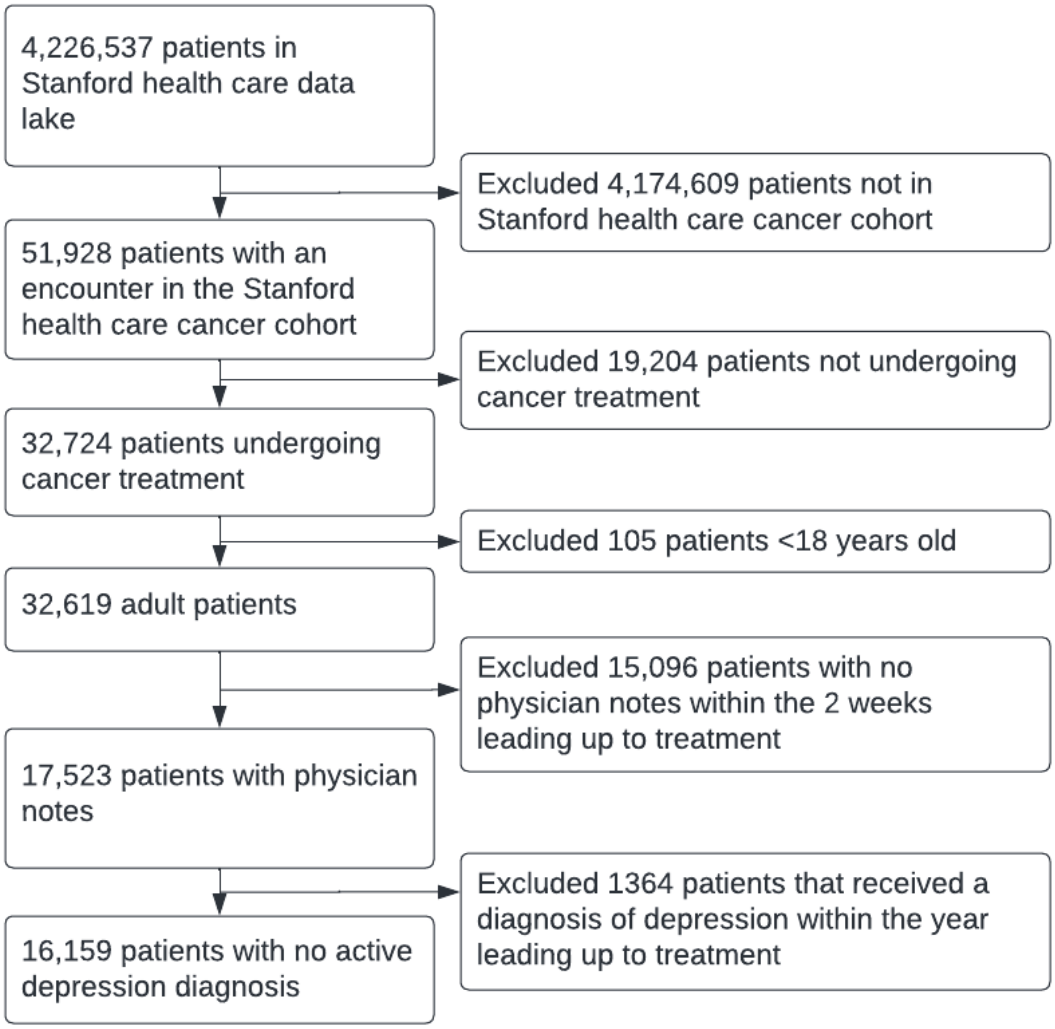
Data Source and Patient Population

This retrospective observational study used data from the integration of 3 health care organizations: an academic medical center (AMC), a primary and specialty care alliance (PSC), and a community medical center (CMC). These organizations offer a wide spectrum of specialized and advanced health care services for complex medical conditions, operating in over 600 clinics. The PSC, established in 2011, comprises more than 70 primary and specialty clinics throughout the California Bay Area. The CMC provides a range of inpatient and outpatient services in the Tri-Valley region of East Bay and was acquired by the AMC in 2015. Following the merger and acquisition, all health care settings adopted the same Epic-based electronic health record (EHR; Epic Systems Corporation) system. Patients for the study were identified from a clinical data warehouse that consolidated patient data from the AMC, PSC, and CMC from 2008 to 2021 [23]. The EHR system was initiated in 2005, and by 2008, the data had reached a state of robustness and high quality. The study concluded in 2021 to ensure that all patients who visited the clinic during the extended period were comprehensively captured.

As an integral component of the EHR system, the MyHealth portal and web interface are seamlessly incorporated into the EHR. This integration includes a patient portal, enabling patients to engage with their health care teams through secure email communication. Patient-generated emails were systematically gathered from the MyHealth patient portal. These email exchanges feature structured subject lines, with patients selecting from a predefined set of categories such as "Non-Urgent Medical Question," "Prescription Question," "Visit Follow-Up Question," "Test Results Question," "Update My Health Information," "Scheduling Question," and "Ordered Test Question." The email body allows for free-text input but is limited to 1000 characters. Importantly, all incoming emails are meticulously triaged to the appropriate members of the patient's health care team, including clerical, scheduling, clinical, or other team members, who take the necessary actions or provide responses as needed [24].

Adult patients receiving chemo- or radiotherapy treatment were included in the cohort. Given the data-intensive nature of the techniques used [25], our objective was to encompass all eligible patients throughout the entire available period at the time of our analysis. The start of cancer treatment was defined as the first patient encounter that registered chemotherapy (including targeted and immunotherapy) or radiotherapy ("chemotherapy" and "clinical procedure codes" in Multimedia Appendix 1). We excluded patients who did not receive cancer treatment (eg, patients seen for a second opinion only), were younger than 18 years, and had no clinician notes within the 2 weeks leading up to the treatment (Figure 1). We also excluded patients with a depression diagnosis within the year leading up to treatment as we aimed to focus on individuals who are at risk of developing depression during or after their treatment (Figure 1). It was assumed that these patients were already receiving treatment for their depression or at least had additional support offered to them.

Figure 1. Flowchart of cohort selection.



Ethical Considerations

This study was approved by the Stanford institutional review board (#47644). Informed consent was waived for this retrospective study for access to personally identifiable health information as it would not be reasonable, feasible, or practical. The data are housed in the Stanford Nero Computing Platform, which is a highly secure, fully integrated internal research data platform meeting all security standards for high risk and protected health information data. The security is managed and monitored, and the platform is updated and adapted to meet regulatory changes.

Predictive Outcome

Depression was defined in consultation with oncologist coauthors (DWB and MR) as a depression diagnosis via the *International Classification of Diseases* (ICD)-9 and ICD-10 codes obtained from EHR data ("ICD depression codes" in Multimedia Appendix 1). This end point was chosen as it was the most conservative and has been shown to correlate reasonably well with clinical opinion [26]. Depression risk was predicted within 1 month of cancer treatment. This time window was chosen as depression prevalence is highest during diagnosis and the acute phase of cancer treatment [27].

Structured Data Predictors

The following variables were obtained from structured EHR fields: sex (male and female), age, insurance status (private, Medicare, Medicaid, and other or not identified), cancer stage (I, II, III, IV, and missing), hospitalized in the previous month (yes or no), 1 or more emergency department visits in the previous month (yes or no), the Charlson comorbidity score [28], and the number of emails sent in the month prior to treatment (none, 1-3, 4, or more) based on a previous study [24]. Insurance status was recoded into 4 comprehensive categories (private, Medicare, Medicaid, and other or not identified). Cancer stage was also recoded to contain the 4 main stages (I, II, III, IV, and missing). Whether or not patients sent emails at night in the previous month was also included as insomnia and depression are intimately related [29]. Binary variables were added indicating whether a patient had previously received a depression diagnosis; depressant medication; or a referral to a psychiatrist, psychologist, or social worker. Finally, race and ethnicity (Hispanic, non-Hispanic Asian, non-Hispanic Black, and non-Hispanic White) was included in one of the sensitivity analyses (see below). The ethnicities "Latino" and "Hispanic" were merged into 1 category (Hispanic). The categorical predictors were converted into dummy variables.

Descriptive statistics were reported in terms of percentages for categorical variables and the mean and SD for continuous variables. We analyzed the cancer and insurance information that was closest to, but preceding, the patient's start of treatment. We stratified descriptive statistics according to outcome (depression diagnosis or not) and messaging behavior (active email communicator in the past month or not).

Unstructured Text Predictors

Unstructured text included patient emails with the subject "Non-Urgent Medical Question" sent through a secure patient portal and clinician notes [24].

A Bidirectional Representations from Transformers (BERT) model was trained on a subset of manually labeled emails to classify each email as being "concerning for depression" or not (see the Multimedia Appendix 2 [30-33] for further details on the annotation strategy and model development). Automatically sent emails; copies of previously sent emails; and emails containing questionnaires, appointment requests, and medication refill requests were removed from the set of patient emails. Emails with less than 30 words were removed from the data set. Each email in the final data set was truncated to a maximum token length of 512. This BERT model assigned each patient email a classification score ranging from 0 (not concerning for depression at all) to 1 (most concerning for depression). These email classification scores were summarized at the patient level by calculating the minimum email classification score in the previous month, the maximum score in the previous month, and the mean score in the previous month. These email classification features were then included as structured data in the subsequent model developments.

Clinician notes that were shorter than 100 words or longer than 5000 were removed as these contained erroneous entries or long copies of previous notes, respectively. Notes with mentions of clinical trials, duplicates, and empty notes were also removed. We merged the most recent clinical notes (at most 3) created within the 2 weeks before the start of treatment. The merged notes were decomposed into chunks of at most 25 sequences (to avoid computational issues), each sequence consisting of 256 tokens.

Model Development

For all models, data were randomly split into the same two-thirds for the train set and one-third for the test set. A total of 6 models were trained to assess the value of multimodal data for this use case.

First, a machine learning (ML) model was developed based on the structured EHR data (model 1), email classification scores (model 2), and the combination of the 2 (model 3). The following ML algorithms were compared for these models: least absolute shrinkage and selection operator (LASSO) logistic regression, a decision tree, random forest, gradient boosting decision trees, *k*-nearest neighbor, and naive Bayes.

LASSO logistic regression is a regularized regression approach, providing both variable selection and shrinkage of regression coefficients. A decision tree is a nonparametric algorithm consisting of a hierarchical tree structure. A random forest combines the predictions of many independently built decision trees into 1 prediction. Gradient boosting decision trees essentially optimize random forest estimation by gradient boosting. The *k*-nearest neighbor algorithm is also nonparametric and uses proximity to previously seen data points to make predictions. Finally, naive Bayes is a generative algorithm that models the distribution of its predictors to make predictions.

The hyperparameters of these models (see Tables S1-S3 in Multimedia Appendix 2) were optimized using Bayesian optimization and 5×10-fold cross-validation. The final ML models were trained on all training data with optimized hyperparameters. The best-performing ML algorithms were the basis for extension with unstructured data.

We trained BERT models based on the clinician notes (model 4), the structured EHR data in combination with the clinician notes (model 5), and the structured EHR data in combination with the email classification scores and the clinician notes (model 6). BERT models are deep learning language models that learn contextual relations between words in a text. Models 5 and 6 made use of a modality-specific deep learning architecture to combine the different data modalities in the modeling process (see Multimedia Appendix 2 for more details) [34]. We used a pretrained DistilBERT model [32] as it required less computation than BERT or ClinicalBERT models [33]. The hyperparameters were tuned on 80% and validated on 20% of the training data. The model parameters of the best-performing epoch on the validation data were chosen for further analyses. Probability estimates were recalibrated via isotonic regression for all models [35].

Statistical Analysis

Model discrimination was quantified by the area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) on the test data. Calibration was assessed through calibration plots, with a calibration intercept and slope as summary performance measures [36]. CIs were obtained via bootstrapping (based on 1000 iterations).

As an initial assessment of clinical usefulness, we performed a decision curve analysis for all 6 models plotting net benefit (NB) across a range of decision probability thresholds [37,38]. NB is defined as the number of true-positive classifications penalized for false-positive classifications [39]. The models have the potential to improve clinical decision-making when they have higher NB than 2 baseline strategies: label all as high risk for developing depression and label none as high risk for developing depression.

Sensitivity Analysis

Sensitivity analyses were performed on the best-performing model to evaluate the impact of modeling choices on model outcomes. Additional models considered different prediction windows (45 days, 2 months, 3 months, and 6 months after the start of cancer treatment). Moreover, patients dying within these prediction windows are a potential competing risk for patients at risk for depression. We therefore removed these patients from the train and test data and repeated the analyses. Variants of outcome definitions such as predicting a prescription of

antidepressant medication and a referral to a psychiatrist, psychologist, or social worker within 1 month of cancer treatment ("antidepressant medication" and "mental health referral" in Multimedia Appendix 1) were considered. These definitions were chosen as they might indicate a patient experiencing depression without being officially diagnosed. We also trained a model on the combined outcome of either receiving a depression diagnosis, antidepressant medication prescription, or referral to a psychiatrist, psychologist, or social worker.

Fairness Analysis and Including Race and Ethnicity

To identify potential fairness issues for specific demographic groups, AUROC, calibration slope, and intercept were compared across sex and race and ethnicity groups [40]. In addition, race and ethnicity was added as a confounder to assess its effect on subgroup model performance.

Software, Data, and Reporting

All analyses were performed in Python 3.9.7 (Python Software Foundation). Code is available in a git repository [41]. We followed the MINIMAR reporting guidelines (see Multimedia Appendix 2) [42].

Results

Descriptive Statistics

A total of 16,159 patients starting cancer treatment between 2008 and 2021 were included in the analyses, of whom 437 (2.7%) received a diagnosis of depression within 1 month of cancer treatment (Table 1 and Figure 1). The 437 patients receiving a depression diagnosis within 1 month of treatment were, on average, younger, more likely to be female, more likely to be non-Hispanic White, and less likely to be non-Hispanic Asian (Table 1). Moreover, patients with a depression diagnosis made more emergency department visits (Table 1). They were also more likely to have received a previous depression diagnosis more than a year before the start of treatment, a prescription for antidepressant medication, and a mental health referral.

Patients who sent emails (4816/16,159, 29.8%) were more likely to be non-Hispanic White or Asian and be privately insured (Table S4 in Multimedia Appendix 2). On average, they were less likely to be hospitalized but made more emergency department visits 1 month prior to treatment and had a higher Charlson comorbidity score; they were also more likely to have previously received a depression diagnosis, antidepressant medication, and a mental health referral.

Table 1. Descriptive statistics of the cancer cohort.

Descriptive statistics	All (N=16,159)	No depression diagnosis within 1 month after onset of treatment (n=15,722, 97.3%)	Depression diagnosis within 1 month after onset of treatment (n=437, 2.7%)
Demographics			
Sex (female), n (%)	8568 (53)	8296 (52.8)[a]	272 (62.2)[a]
Age (years), mean (SD)	62 (15)	62 (15)[a]	60 (14)[a]
Race and ethnicity, n (%)			
Hispanic	1870 (11.6)	1812 (11.5)[a]	58 (13.3)[a]
Non-Hispanic Asian	3582 (22.2)	3525 (22.4)[a]	57 (13)[a]
Non-Hispanic Black	422 (2.6)	410 (2.6)[a]	<20 (<5)[a]
Non-Hispanic White	8864 (54.9)	8583 (54.6)[a]	281 (64.3)[a]
Other	1421 (8.8)	1392 (8.9)[a]	29 (6.6)[a]
Insurance characteristics, n (%)			
Private	8745 (54.1)	8496 (54)	249 (57)
Medicare	2590 (16)	2514 (16)	76 (17.4)
Medicaid	1917 (11.9)	1860 (11.8)	57 (13)
Other or not identified	2907 (18)	2852 (18.1)	55 (12.6)
Treatment characteristics, mean (SD)			
Number of hospitalizations one month prior to treatment	2083 (13)	2016 (13)	67 (15)
Number of emergency department visits 1 month prior to treatment	945 (6)	895 (6)[a]	50 (11)[a]
Charlson comorbidity score	6.9 (3.8)	6.9 (3.8)	6.9 (3.9)
Tumor type, n (%)			
Breast	1772 (11)	1739 (11.1)	33 (7.6)
Lung	1001 (6.2)	973 (6.2)	28 (6.4)
Prostate	777 (4.8)	764 (4.9)	<20 (<5)
Colon and rectum	543 (3.4)	525 (3.3)	<20 (<5)
Non-Hodgkin lymphoma	535 (3.3)	527 (3.4)	<20 (<5)
Other	3459 (21.4)	3364 (21.4)	95 (21.7)
Missing	8072 (50)	7830 (49.8)	242 (55.4)
Cancer stage, n (%)			
Stage I	1492 (9.2)	1466 (9.3)	26 (5.9)
Stage II	1499 (9.3)	1468 (9.3)	31 (7.1)
Stage III	1329 (8.2)	1294 (8.2)	35 (8)
Stage IV	1758 (10.9)	1699 (10.8)	59 (13.5)
Missing	10081 (62.4)	9795 (62.3)	286 (65.4)
Patient email information (1 month prior to treatment)			
Sent 1 or more emails, n (%)	4070 (25.2)	3943 (25.1)	127 (29.1)
Email length in words, mean (SD)	49 (35)	49 (35)	49 (35)
Sent emails at night, n (%)	308 (1.9)	296 (1.9)	<20 (<5)
Mental health history, n (%)			
History of depression diagnosis	400 (2.5)	343 (2.2)[a]	57 (13)[a]

Descriptive statistics	All (N=16,159)	No depression diagnosis within 1 month after onset of treatment (n=15,722, 97.3%)	Depression diagnosis within 1 month after onset of treatment (n=437, 2.7%)
History of antidepressant medication	2219 (13.7)	2030 (12.9)[a]	189 (43.2)[a]
History of mental health referral	2707 (16.8)	2563 (16.3)[a]	144 (33)[a]

[a]This was tested at the 5% significance level.

Performance Statistics

The best-performing ML models were based on LASSO logistic regression (Table 2; Tables S1 and S3 in Multimedia Appendix 2). The model based on structured data alone had an AUROC of 0.74 (95% CI 0.71-0.78). The combination of structured data with email classification scores also had an AUROC of 0.74 (95% CI 0.71-0.78), while a model based solely on email classification scores had an AUROC of 0.54 (95% CI 0.52-0.56). At a high level of sensitivity (0.9 at a decision threshold of 1%; Table 3), the PPV of the best-performing model based on structured data was low (0.04; Table 3). At higher decision thresholds (3% and 10%; Table 3), the PPV was increased to 0.07 and 0.17, respectively, but this came at a cost of sensitivity (0.63 and 0.19).

The BERT model based on the clinician notes performed worst and had an AUROC of 0.50 (95% CI 0.49-0.52; Table 2). Combining structured EHR data with clinician notes did improve AUROC performance (0.71, 95% CI 0.68-0.75; Table 2) and so did adding email classification scores (0.70, 95% CI 0.67-0.73; Table 2).

Calibration was acceptable for all ML models. The BERT-based models tended to produce overly extreme risk estimates even after recalibration.

The decision curve analysis showed a small range of decision thresholds for which the best-performing model (LASSO logistic regression based on structured data) had higher NB than the treat all or treat no one strategies (Figure 2). At a decision threshold of 3%, the model with structured EHR data had a NB of 0.01. This represents a net increase of 1 true positive patient at risk for depression per 100 patients without increasing any false positives (at the start of treatment). At a threshold of 10%, the model had a NB of only 0.002, so 2 net true positives per 1000 patients.

Table 2. Discrimination and calibration for predicting depression risk within 1 month after the onset of treatment (test data).

Type of data	AUROC[a] (95% CI)	Calibration intercept (95% CI)	Calibration slope (95% CI)
Structured EHR[b] data	0.74 (0.71 to 0.78)	0.07 (–0.09 to 0.24)	0.93 (0.77 to 1.09)
Patient emails	0.54 (0.52 to 0.56)	–0.02 (–0.18 to 0.14)	1.0 (0.52 to 1.48)
Structured EHR data and patient emails	0.74 (0.71 to 0.78)	0.07 (–0.09 to 0.24)	0.91 (0.76 to 1.07)
Clinician notes	0.5 (0.49 to 0.52)	–0.05 (–0.21 to 0.11)	0.94 (–1.32 to 3.2)
Structured EHR data and clinician notes	0.71 (0.68 to 0.75)	–0.09 (–0.25 to 0.07)	1.92 (1.57 to 2.28)
Structured EHR data, clinician notes, and patient emails	0.7 (0.67 to 0.73)	–0.16 (–0.32 to –0.0)	2.46 (1.98 to 2.93)

[a]AUROC: area under the receiver operating characteristics curve.

[b]EHR: electronic health record.

Table 3. Sensitivity, specificity, PPV[a], and NPV[b] at different decision thresholds for predicting depression risk within 1 month after the onset of treatment (test data).

Threshold and analysis	Structured EHR[c] data	Patient emails	Structured EHR data and patient emails	Clinician notes	Structured EHR data and clinician notes	Structured EHR data, clinician notes, and patient emails
1%						
Sensitivity (n/N)	0.9 (140/156)	1.0 (156/156)	0.87 (136/156)	1.0 (156/156)	1.0 (156/156)	1.0 (156/156)
Specificity (n/N)	0.35 (1847/5231)	0.0 (0/5231)	0.37 (1915/5231)	0.0 (0/5231)	0.0 (0/5231)	0.0 (0/5231)
PPV (n/N)	0.04 (140/3524)	0.03 (156/5387)	0.04 (136/3452)	0.03 (156/5387)	0.03 (156/5387)	0.03 (156/5387)
NPV (n/N)	0.99 (1847/1863)	N/A[d]	0.99 (1915/1935)	N/A	N/A	N/A
3%						
Sensitivity (n/N)	0.63 (98/156)	0.13 (20/156)	0.58 (90/156)	1.0 (156/156)	0.55 (86/156)	0.67 (104/156)
Specificity (n/N)	0.75 (3912/5231)	0.95 (4962/5231)	0.77 (4032/5231)	0.0 (0/5231)	0.82 (4293/5231)	0.71 (3735/5231)
PPV (n/N)	0.07 (98/1417)	0.07 (20/289)	0.07 (90/1289)	0.03 (156/5387)	0.08 (86/1024)	0.06 (104/1600)
NPV (n/N)	0.99 (3912/3970)	0.97 (4962/5098)	0.98 (4032/4098)	N/A	0.98 (4293/4363)	0.99 (3735/3787)
10%						
Sensitivity (n/N)	0.19 (29/156)	0.0 (0/156)	0.19 (30/156)	0.0 (0/156)	0.0 (0/156)	0.0 (0/156)
Specificity (n/N)	0.97 (5086/5231)	1.0 (5231/5231)	0.97 (5071/5231)	1.0 (5231/5231)	1.0 (5231/5231)	1.0 (5231/5231)
PPV (n/N)	0.17 (29/174)	N/A	0.16 (30/190)	N/A	N/A	N/A
NPV (n/N)	0.98 (5086/5213)	0.97 (5231/5387)	0.98 (5071/5197)	0.97 (5231/5387)	0.97 (5231/5387)	0.97 (5231/5387)

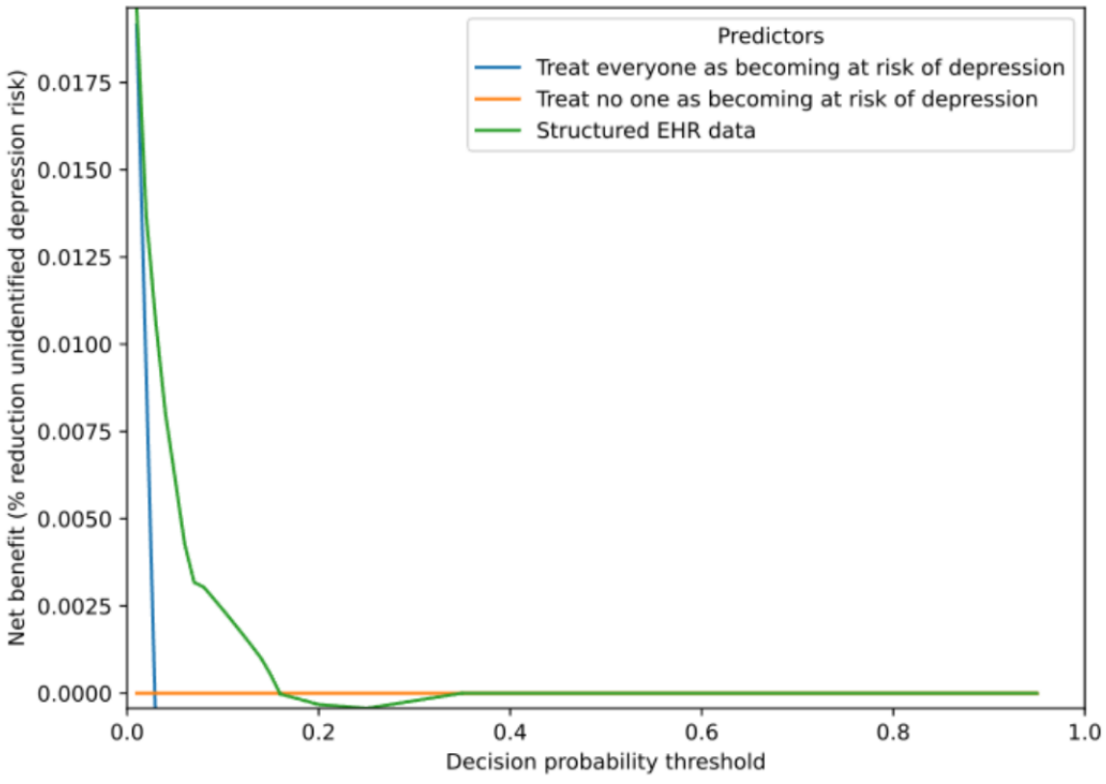
[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]EHR: electronic health record.

[d]N/A: not available.

Figure 2. This decision curve analysis (DCA) plots net benefit for the baseline treat all and treat none strategies and the best-performing prediction model (LASSO logistic regression on structured data). EHR: electronic health record; LASSO: least absolute shrinkage and selection operator.



Sensitivity Analysis

The models predicting depression risk within 45 days (AUROC 0.73, 95% CI 0.69-0.76), 2 months (AUROC 0.73, 95% CI 0.70-0.76), 3 months (AUROC 0.73, 95% CI 0.71-0.76), and 6 months (AUROC 0.72, 95% CI 0.70-0.74) of cancer treatment obtained similar discrimination and calibration compared to the base model predicting depression risk within 1 month (LASSO logistic regression; Table S5 in Multimedia Appendix 2). In the test data, a total of 24 (0.4%) patients died within 1 month after starting treatment. Omitting patients dying within the time frames of interest (1 month-6 months) had no impact on model performance (Table S6 in Multimedia Appendix 2). The model trained to predict depression medication (LASSO logistic regression) also obtained similar discrimination (0.75, 95% CI 0.73-0.78; Table S7 in Multimedia Appendix 2) and calibration compared to the base model predicting depression risk via depression diagnosis. The model trained to predict a referral to a psychiatrist, psychologist, or social worker obtained a lower AUROC of 0.62 (95% CI 0.60-0.64; Table S7 in Multimedia Appendix 2) and comparable calibration.

Fairness Analysis and Including Race and Ethnicity

The fairness analysis showed that model discrimination was similar for male patients (AUROC 0.73, 95% CI 0.67-0.80; Table S8 in Multimedia Appendix 2) and female patients (AUROC 0.74, 95% CI 0.70-0.78; Table S8 in Multimedia Appendix 2). The calibration plot showed that depression risk was underestimated for female patients and overestimated for male patients (Figure S1 in Multimedia Appendix 2). Discrimination was best for the non-Hispanic Black patients (AUROC 0.92, 95% CI 0.84-0.99; Table S8 in Multimedia Appendix 2), with respect to the non-Hispanic White patients (AUROC 0.74, 95% CI 0.69-0.78; Table S8 in Multimedia Appendix 2) and the non-Hispanic Asian patients (AUROC 0.75, 95% CI 0.63-0.87; Table S8 in Multimedia Appendix 2), and it was worst for Hispanic patients (AUROC 0.71, 95% CI 0.62-0.80; Table S8 in Multimedia Appendix 2). Predictions were underestimated for the non-Hispanic Black patients and overestimated for the non-Hispanic Asian patients (Figure S1 in Multimedia Appendix 2). Adding race and ethnicity as a feature to the best-performing model did not improve model discrimination or calibration (AUROC 0.74, 95% CI 0.71-0.78 vs 0.74, 95% CI 0.70-0.77; Table S9 in Multimedia Appendix 2).

Discussion

Principal Findings

This study developed a prediction model to identify patients with cancer at risk for depression within 1 month of chemo- or radiotherapy treatment. We used data from a large comprehensive cancer center with over 16,000 patients. The best-performing models (LASSO logistic regression with structured data with or without patient email classification scores) had reasonable AUROC and calibration. The LASSO logistic regression model with structured data demonstrates a small improvement in NB over the baseline strategy of labeling no one as at risk for depression. Multimodal BERT models (trained on structured data and unstructured text) did not perform better than the best-performing ML model trained solely on structured data.

To date, depression in patients with cancer is underdiagnosed, and studies show that patients with depression are up to 3 times more likely to be noncompliant with medical treatment recommendations [3,43,44]. Treatment adherence is a high priority, given the evidence demonstrating statistically significant associations between treatment nonadherence and patient outcomes, including cancer progression, low-value health care use, and worse survival [45-48]. Therefore, an AI model—which flags patients at risk for depression with minimal clinical input and workflow disruption—is needed at the point of care to prompt clinicians to intervene early and improve patient well-being and anticancer outcomes.

This model may be used in preparation for clinical consultations to more efficiently use the limited time allotted to oncologist-patient interaction to facilitate any needed additional mental health support. By harnessing a combination of structured EHR data and unstructured text data from patient emails and clinician notes, the tool can offer a comprehensive assessment of a patient's depression risk and help synthesize this information at point of care for the provider. With the ability to establish personalized risk assessments, determine clinical use thresholds, and address potential biases in risk assessment, a clinical decision support tool developed from this work has the potential to significantly enhance the quality of care and mental health outcomes for these vulnerable patients. As the study recognizes the need for ongoing validation, refinement, and bias mitigation, it underscores the dynamic and adaptable nature of this tool in improving cancer care and treatment adherence. This tool can be a valuable addition to the health care system, ultimately improving mental health outcomes and treatment adherence for these vulnerable patients.

The created model has good performance, although our label (receiving a depression diagnosis) depends heavily upon the accurate recognition of depression by the care team. The model's clinical usefulness depends on the acceptability of the test trade-off. The best-performing model had a high false-positive rate at high levels of sensitivity, and the decision curve analysis showed a test trade-off of 100 assessments for 1 additional true positive patient at a decision threshold of 3%. If these assessments can be done nearly for free (eg, a quick check during a patient visit) and if we already miss all future depressions, then this small improvement may be welcome, although this warrants further validation and testing in the clinical environment. The high false-positive rate and small NB of the best-performing model are likely affected by the moderate discrimination and low event rate [49]. In future developments, the NB may be increased by focusing on improving the labeling of the outcome variable. In addition, richer input data not available to us at the time of analysis could improve model discrimination, like information on lifestyle habits, self-reported mental health assessments, and clinical and pathological factors.

As depression presents differently across sex, race, and ethnicity [50-52], algorithmic fairness forms an important concern when predicting depression risk. We found discrepant model calibration across race, ethnicity, and sex even when controlling

for race, ethnicity, and sex in the model. These results align with previous findings that showed poor calibration for minority groups [53,54] and stress the importance of algorithmic fairness assessment in the depression domain. The differences in calibration may be caused by different (recorded) depression rates among groups. This could result in a disproportionate number of missed patients in need of additional mental health resources in specific groups. For example, female and non-Hispanic Black patients might consistently receive a lower predicted risk score than their actual risk. A next step could be to apply bias mitigation techniques for in- or postprocessing during model development, like threshold selection and recalibration within specific groups [55]. Moreover, more diverse data may be collected to adequately capture the differences in symptomatology between different groups. For example, we may include appetite disturbances that are reported more by women and comorbid alcohol and substance abuse that are reported more by men [50].

We also found discrepant model discrimination across race and ethnicity, with the highest AUROC for the non-Hispanic Black group. These findings diverge from the literature, where the AUROC of the minority groups is usually lower compared to the majority group [56]. However, caution is needed when interpreting this finding, due to the very low number of positive cases in this group (less than 20). More data should be collected to better investigate these differences.

The models based solely on text information (patient emails and clinician notes) performed on par with a random coin toss. This implies that the signal-to-noise ratio in this type of data may be too low to be of prognostic value for this specific use case. This might be particularly true for patient emails, where the frequency of the emails varied widely between patients. However, it is important to note that unstructured text, such as patient emails and clinician notes, can potentially provide valuable information that is not captured in structured data. Therefore, multimodal models that incorporate both structured and unstructured data have the potential to improve clinical predictions. Increasing and regularizing the frequency of digital contact between patient and clinician may aid future research on multimodal models in this field, for example, through digital systems for monitoring patient-reported outcomes [57,58]. Digital communication with the aid of chat robots such as ChatGPT [59] provides further direction to better capture patients' mental health status. This finding also implies that structured data contains strong predictors for depression risk, for example, a history of depression or mental illness, which is well established in the literature and should be considered for future model developments [60-62].

Limitations

This study had limitations. First, we used the ICD codes for depression diagnosis as indicators of depression risk. This provided a clear and detectable label for our outcome event in the EHR. However, not all patients experiencing depression will receive a coded depression diagnosis with a related ICD code as underdiagnosis is a common problem [3,9]. It is possible that depression may have been diagnosed elsewhere and not recorded in our EHR, that depressive symptoms may have existed and not been recorded or ignored by the oncology-focused clinicians, or that the patient did not express their depressive symptoms to their oncology-focused clinician. In addition, some inconsistencies persisted between the ICD-9 and ICD-10 codes, with the ICD-10 codes including depression associated with bipolar disorder. This may have compromised the accuracy of our predictive models in this exploratory study and should be considered for future research.

Moreover, changing the outcome of interest to either antidepressant medication or a referral to a psychiatrist, psychologist, or social worker did not change the accuracy of the predictive models. An explanation might be that patients with depression are often treated with antidepressants by primary care doctors. For antidepressant medication, it is important to note that there may have been overascertainment as this medication is also used to treat more severe and chronic forms of anxiety. This should be considered when interpreting our results and warrants further study.

Second, the modeling approach was focused on a point-of-care solution, meaning we used clinically meaningful end points (eg, 1 month after starting cancer treatment) and used a diverse patient population. Although this provides the potential for broad application across multiple cancer types, the diversity in cancer types and cancer stages might have introduced noise and impacted model performance.

Third, we used cut-off values for clinician notes that were too short or too long to keep the modeling computationally feasible. This may have led to information loss. Future research may investigate ways of retaining this information when preprocessing texts. Finally, we used data from a single integrated health system for model development, albeit comprised of 3 sites (academic hospital, community hospital, and community practice network). As the cultural background of patients and some data are specific to this health system, our results may not generalize to other populations. Further validation on data sets with different demographics and examination of the mechanisms driving potential biases are needed.

Conclusions

This study demonstrated the potential and limitations of using structured and unstructured text data for predicting depression risk in patients with cancer using a variety of ML and multimodal models. After further validation and mitigating biases across subgroups, these models have the potential to improve patient outcomes by alerting clinicians of the possible need to escalate support among this vulnerable patient population. Future studies might improve the prediction of depression risk in patients with cancer by refining the outcome label, expanding the predictors related to mental health, and devoting part of the digital patient communication to mental health aspects.

Acknowledgments

We like to thank Max Schuessler, Vaibhavi Shah, and Angelo Capodici for their help with the annotations of patient emails. Our special thanks go to Dr David Spiegel for reviewing this article for psychiatric relevance and accuracy. This work was funded by the Leids Universiteits Fonds/Slingelands Fonds, the Prins Bernhard Cultuurfonds or Crone-Haver Droeze Fonds, and Fonds Dr Catharine van Tussenbroek. These funders played no role in study design, data collection, analysis, and interpretation of data, or writing the manuscript.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to the protected nature of the patient data. Requests to access these data sets should be directed to boussard@stanford.edu.

Authors' Contributions

AdH, MvB, and THB were responsible for the conceptualization and design of the study. AdH performed the data extraction. AdH and CF performed the data analysis. MR and DB provided clinical advice and recommendations on usability and clinical relevance. AdH drafted the original manuscript. All authors had full access to all the data, critically analyzed, reviewed, contributed, and approved the final manuscript.

Conflicts of Interest

DB reports institutional research funding from BeyondSpring, leadership roles or stock ownership in Artelo and Madora, and personal fees from G1 Therapeutics, Bristol Myers Squibb, Merck & Co Inc, and Eli Lilly and Company all outside the submitted work. THB is a board member and stockholder of Athelo Health, a stockholder at Verantos, Inc, and a consultant for Grai-Matter outside the submitted work. The other authors declare no competing interests.

Multimedia Appendix 1
Tabular metadata appendix.
[XLS File (Microsoft Excel File), 47 KB - medinform_v12i1e51925_app1.xls]

Multimedia Appendix 2
Supplemental methods and results.
[DOCX File , 117 KB - medinform_v12i1e51925_app2.docx]

References

1. Linden W, Vodermaier A, Mackenzie R, Greig D. Anxiety and depression after cancer diagnosis: prevalence rates by cancer type, gender, and age. J Affect Disord 2012;141(2-3):343-351. [doi: 10.1016/j.jad.2012.03.025] [Medline: 22727334]
2. Smith HR. Depression in cancer patients: pathogenesis, implications and treatment (review). Oncol Lett 2015;9(4):1509-1514 [FREE Full text] [doi: 10.3892/ol.2015.2944] [Medline: 25788991]
3. Pitman A, Suleman S, Hyde N, Hodgkiss A. Depression and anxiety in patients with cancer. BMJ 2018;361:k1415 [FREE Full text] [doi: 10.1136/bmj.k1415] [Medline: 29695476]
4. Colleoni M, Mandala M, Peruzzotti G, Robertson C, Bredart A, Goldhirsch A. Depression and degree of acceptance of adjuvant cytotoxic drugs. Lancet 2000;356(9238):1326-1327. [doi: 10.1016/S0140-6736(00)02821-X] [Medline: 11073026]
5. Grassi L, Indelli M, Marzola M, Maestri A, Santini A, Piva E, et al. Depressive symptoms and quality of life in home-care-assisted cancer patients. J Pain Symptom Manage 1996;12(5):300-307 [FREE Full text] [doi: 10.1016/s0885-3924(96)00181-9] [Medline: 8942125]
6. National Survey on Drug Use and Health (NSDUH). Substance Abuse and Mental Health Services Administration. 2020. URL: https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health [accessed 2023-12-15]
7. Walker J, Hansen CH, Martin P, Symeonides S, Ramessur R, Murray G, et al. Prevalence, associations, and adequacy of treatment of major depression in patients with cancer: a cross-sectional analysis of routinely collected clinical data. Lancet Psychiatry 2014;1(5):343-350 [FREE Full text] [doi: 10.1016/S2215-0366(14)70313-X] [Medline: 26360998]
8. Caruso R, Breitbart W. Mental health care in oncology. Contemporary perspective on the psychosocial burden of cancer and evidence-based interventions. Epidemiol Psychiatr Sci 2020;29:e86 [FREE Full text] [doi: 10.1017/S2045796019000866] [Medline: 31915100]
9. Mitchell AJ, Chan M, Bhatti H, Halton M, Grassi L, Johansen C, et al. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies. Lancet Oncol 2011;12(2):160-174. [doi: 10.1016/S1470-2045(11)70002-X] [Medline: 21251875]
10. Vinckx MA, Bossuyt I, de Casterlé BD. Understanding the complexity of working under time pressure in oncology nursing: a grounded theory study. Int J Nurs Stud 2018;87:60-68. [doi: 10.1016/j.ijnurstu.2018.07.010] [Medline: 30055374]

11. Dreismann L, Goretzki A, Ginger V, Zimmermann T. What if… I asked cancer patients about psychological distress? barriers in psycho-oncological screening from the perspective of nurses-a qualitative analysis. Front Psychiatry 2021;12:786691 [FREE Full text] [doi: 10.3389/fpsyt.2021.786691] [Medline: 35153856]
12. Söllner W, DeVries A, Steixner E, Lukas P, Sprinzl G, Rumpold G, et al. How successful are oncologists in identifying patient distress, perceived social support, and need for psychosocial counselling? Br J Cancer 2001;84(2):179-185 [FREE Full text] [doi: 10.1054/bjoc.2000.1545] [Medline: 11161373]
13. Steven B, Lange L, Schulz H, Bleich C. Views of psycho-oncologists, physicians, and nurses on cancer care-a qualitative study. PLoS One 2019;14(1):e0210325 [FREE Full text] [doi: 10.1371/journal.pone.0210325] [Medline: 30650112]
14. 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;21(11):116. [doi: 10.1007/s11920-019-1094-0]
15. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep 2021;11(1):1980 [FREE Full text] [doi: 10.1038/s41598-021-81368-4] [Medline: 33479383]
16. Arun V, Prajwal V, Krishna M, Arunkumar BV, Padma SK, Shyam V. A boosted machine learning approach for detection of depression. 2018 Presented at: 2018 IEEE Symposium Series on Computational Intelligence (SSCI); November 18-21, 2018; Bangalore, India. [doi: 10.1109/ssci.2018.8628945]
17. Patel MJ, Andreescu C, Price JC, Edelman KL, Reynolds CF, Aizenstein HJ. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. Int J Geriatr Psychiatry 2015;30(10):1056-1067 [FREE Full text] [doi: 10.1002/gps.4262] [Medline: 25689482]
18. Aldarwish MM, Ahmad HF. Predicting depression levels using social media posts. 2017 Presented at: 2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS); March 22-24, 2017; Bangkok, Thailand. [doi: 10.1109/isads.2017.41]
19. Deshpande M, Rao V. Depression detection using emotion artificial intelligence. 2017 Presented at: 2017 International Conference on Intelligent Sustainable Systems (ICISS); December 07-08, 2017; Palladam, India. [doi: 10.1109/iss1.2017.8389299]
20. Ricard BJ, Marsch LA, Crosier B, Hassanpour S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. J Med Internet Res 2018;20(12):e11817 [FREE Full text] [doi: 10.2196/11817] [Medline: 30522991]
21. Papachristou N, Puschmann D, Barnaghi P, Cooper B, Hu X, Maguire R, et al. Learning from data to predict future symptoms of oncology patients. PLoS One 2018;13(12):e0208808 [FREE Full text] [doi: 10.1371/journal.pone.0208808] [Medline: 30596658]
22. Chen L, Ma X, Zhu N, Xue H, Zeng H, Chen H, et al. Facial expression recognition with machine learning and assessment of distress in patients with cancer. Oncol Nurs Forum 2021;48(1):81-93 [FREE Full text] [doi: 10.1188/21.ONF.81-93] [Medline: 33337433]
23. Sun R, Bozkurt S, Winget M, Cullen MR, Seto T, Hernandez-Boussard T. Characterizing patient flow after an academic hospital merger and acquisition. Am J Manag Care 2021;27(10):e343-e348 [FREE Full text] [doi: 10.37765/ajmc.2021.88764] [Medline: 34668676]
24. Coquet J, Blayney DW, Brooks JD, Hernandez-Boussard T. Association between patient-initiated emails and overall 2-year survival in cancer patients undergoing chemotherapy: evidence from the real-world setting. Cancer Med 2020;9(22):8552-8561 [FREE Full text] [doi: 10.1002/cam4.3483] [Medline: 32986931]
25. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14(1):137 [FREE Full text] [doi: 10.1186/1471-2288-14-137] [Medline: 25532820]
26. Trinh NHT, Youn SJ, Sousa J, Regan S, Bedoya CA, Chang TE, et al. Using electronic medical records to determine the diagnosis of clinical depression. Int J Med Inform 2011;80(7):533-540 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.03.014] [Medline: 21514880]
27. Krebber AMH, Buffart LM, Kleijn G, Riepma IC, de Bree R, Leemans CR, et al. Prevalence of depression in cancer patients: a meta-analysis of diagnostic interviews and self-report instruments. Psychooncology 2014;23(2):121-130 [FREE Full text] [doi: 10.1002/pon.3409] [Medline: 24105788]
28. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383 [FREE Full text] [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
29. Benca RM, Peterson MJ. Insomnia and depression. Sleep Med 2008;9(Suppl 1):S3-S9 [FREE Full text] [doi: 10.1016/S1389-9457(08)70010-8] [Medline: 18929317]
30. Zhang. Improved Adam optimizer for deep neural networks. 2018 Presented at: 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS); June 4-6, 2018; Banff, AB, Canada.
31. Lamproudis, Henriksson, Dalianis. Developing a clinical language model for Swedish: continued pretraining of generic BERT with in-domain data. In: Recent Advances in Natural Language Processing. Kerrville, TX: Association for

Computational Linguistics; 2021 Presented at: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021); September 1-3, 2021; Held Online.
32. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019 Presented at: 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019; December 13, 2019; Vancouver, BC, Canada. [doi: 10.1090/mbk/121/79]
33. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019 Presented at: 2nd Clinical Natural Language Processing (ClinicalNLP) Workshop at NAACL 2019; June 7, 2019; Minneapolis, USA p. 72-78. [doi: 10.18653/v1/w19-1909]
34. Fanconi C, van Buchem M, Hernandez-Boussard T. Natural language processing methods to identify oncology patients at high risk for acute care with clinical notes. AMIA Jt Summits Transl Sci Proc 2023;2023:138-147 [FREE Full text] [Medline: 37350895]
35. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. ICML 2001;1:609-616 [FREE Full text]
36. van Calster B, Nieboer D, Vergouwe Y, de Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167-176 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.12.005] [Medline: 26772608]
37. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health 2022;4(12):e853-e855 [FREE Full text] [doi: 10.1016/S2589-7500(22)00188-1] [Medline: 36270955]
38. Vickers AJ, van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6 [FREE Full text] [doi: 10.1136/bmj.i6] [Medline: 26810254]
39. 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;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
40. Röösli E, Bozkurt S, Hernandez-Boussard T. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. Sci Data 2022;9(1):24 [FREE Full text] [doi: 10.1038/s41597-021-01110-7] [Medline: 35075160]
41. Predicting depression for cancer patients. gitlab. URL: https://gitlab.com/a.a.h.de_hond/predicting-depression-for-cancer-patients [accessed 2024-01-04]
42. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (Minimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020;27(12):2011-2015 [FREE Full text] [doi: 10.1093/jamia/ocaa088] [Medline: 32594179]
43. DiMatteo MR, Lepper HS, Croghan TW. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. Arch Intern Med 2000;160(14):2101-2107 [FREE Full text] [doi: 10.1001/archinte.160.14.2101] [Medline: 10904452]
44. Gold SM, Köhler-Forsberg O, Moss-Morris R, Mehnert A, Miranda JJ, Bullinger M, et al. Comorbid depression in medical diseases. Nat Rev Dis Primers 2020;6(1):69 [FREE Full text] [doi: 10.1038/s41572-020-0200-2] [Medline: 32820163]
45. Makubate B, Donnan PT, Dewar JA, Thompson AM, McCowan C. Cohort study of adherence to adjuvant endocrine therapy, breast cancer recurrence and mortality. Br J Cancer 2013;108(7):1515-1524 [FREE Full text] [doi: 10.1038/bjc.2013.116] [Medline: 23519057]
46. Wu EQ, Johnson S, Beaulieu N, Arana M, Bollu V, Guo A, et al. Healthcare resource utilization and costs associated with non-adherence to imatinib treatment in chronic myeloid leukemia patients. Curr Med Res Opin 2010;26(1):61-69 [FREE Full text] [doi: 10.1185/03007990903396469] [Medline: 19905880]
47. Hershman DL, Shao T, Kushi LH, Buono D, Tsai WY, Fehrenbacher L, et al. Early discontinuation and non-adherence to adjuvant hormonal therapy are associated with increased mortality in women with breast cancer. Breast Cancer Res Treat 2011;126(2):529-537 [FREE Full text] [doi: 10.1007/s10549-010-1132-4] [Medline: 20803066]
48. Giese-Davis J, Collie K, Rancourt KMS, Neri E, Kraemer HC, Spiegel D. Decrease in depression symptoms is associated with longer survival in patients with metastatic breast cancer: a secondary analysis. J Clin Oncol 2011;29(4):413-420 [FREE Full text] [doi: 10.1200/JCO.2010.28.4455] [Medline: 21149651]
49. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. Lancet Digit Health 2022;4(12):e853-e855 [FREE Full text] [doi: 10.1016/S2589-7500(22)00188-1] [Medline: 36270955]
50. Altemus M, Sarvaiya N, Epperson CN. Sex differences in anxiety and depression clinical perspectives. Front Neuroendocrinol 2014;35(3):320-330 [FREE Full text] [doi: 10.1016/j.yfrne.2014.05.004] [Medline: 24887405]
51. Barnes DM, Keyes KM, Bates LM. Racial differences in depression in the United States: how do subgroup analyses inform a paradox? Soc Psychiatry Psychiatr Epidemiol 2013;48(12):1941-1949 [FREE Full text] [doi: 10.1007/s00127-013-0718-7] [Medline: 23732705]
52. Hooker K, Phibbs S, Irvin VL, Mendez-Luck CA, Doan LN, Li T, et al. Depression among older adults in the United States by disaggregated race and ethnicity. Gerontologist 2019;59(5):886-891 [FREE Full text] [doi: 10.1093/geront/gny159] [Medline: 30561600]

53. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 2021;27(12):2176-2182 [FREE Full text] [doi: 10.1038/s41591-021-01595-0] [Medline: 34893776]
54. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA. Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. Lancet Digit Health 2021;3(4):e241-e249 [FREE Full text] [doi: 10.1016/S2589-7500(21)00022-4] [Medline: 33766288]
55. Pfohl S, Xu Y, Foryciarz A, Ignatiadis N, Genkins J, Shah N. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. 2022 Presented at: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency; June 21-24, 2022; Seoul Republic of Korea. [doi: 10.1145/3531146.3533166]
56. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366(6464):447-453 [FREE Full text] [doi: 10.1126/science.aax2342] [Medline: 31649194]
57. Denis F, Basch E, Septans AL, Bennouna J, Urban T, Dueck AC, et al. Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. JAMA 2019;321(3):306-307 [FREE Full text] [doi: 10.1001/jama.2018.18085] [Medline: 30667494]
58. Basch E, Stover AM, Schrag D, Chung A, Jansen J, Henson S, et al. Clinical utility and user perceptions of a digital system for electronic patient-reported symptom monitoring during routine cancer care: findings from the PRO-TECT trial. JCO Clin Cancer Inform 2020;4:947-957 [FREE Full text] [doi: 10.1200/CCI.20.00081] [Medline: 33112661]
59. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv Preprint posted online on March 4, 2022. [doi: 10.48550/arXiv.2203.02155]
60. Stafford L, Komiti A, Bousman C, Judd F, Gibson P, Mann GB, et al. Predictors of depression and anxiety symptom trajectories in the 24 months following diagnosis of breast or gynaecologic cancer. Breast 2016;26:100-105 [FREE Full text] [doi: 10.1016/j.breast.2016.01.008] [Medline: 27017248]
61. Fervaha G, Izard JP, Tripp DA, Aghel N, Shayegan B, Klotz L, et al. Psychological morbidity associated with prostate cancer: rates and predictors of depression in the RADICAL PC study. Can Urol Assoc J 2021;15(6):181-186 [FREE Full text] [doi: 10.5489/cuaj.6912] [Medline: 33212008]
62. Wojnarowski C, Firth N, Finegan M, Delgadillo J. Predictors of depression relapse and recurrence after cognitive behavioural therapy: a systematic review and meta-analysis. Behav Cogn Psychother 2019;47(5):514-529 [FREE Full text] [doi: 10.1017/S1352465819000080] [Medline: 30894231]

Abbreviations

AI: artificial intelligence
AMC: academic medical center
AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
CMC: community medical center
EHR: electronic health record
ICD: International Classification of Diseases
LASSO: least absolute shrinkage and selection operator
ML: machine learning
NB: net benefit
NPV: negative predictive value
PPV: positive predictive value
PSC: primary and specialty care alliance

Edited by C Lovis; submitted 17.08.23; peer-reviewed by L Liu, Y Chu; comments to author 03.10.23; revised version received 11.11.23; accepted 08.12.23; published 18.01.24.

Please cite as:
de Hond A, van Buchem M, Fanconi C, Roy M, Blayney D, Kant I, Steyerberg E, Hernandez-Boussard T
Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study
JMIR Med Inform 2024;12:e51925
URL: https://medinform.jmir.org/2024/1/e51925
doi:10.2196/51925
PMID:38236635

Original Paper

Improving Prediction of Survival for Extremely Premature Infants Born at 23 to 29 Weeks Gestational Age in the Neonatal Intensive Care Unit: Development and Evaluation of Machine Learning Models

Angie Li[1], MD; Sarah Mullin[1], PhD; Peter L Elkin[1], MD

Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, United States

Corresponding Author:
Angie Li, MD
Department of Biomedical Informatics
Jacobs School of Medicine and Biomedical Sciences
University at Buffalo
77 Goodell Street
Suite 540
Buffalo, NY, 14203
United States
Phone: 1 716 888 4858
Email: ali83@buffalo.edu

Abstract

Background: Infants born at extremely preterm gestational ages are typically admitted to the neonatal intensive care unit (NICU) after initial resuscitation. The subsequent hospital course can be highly variable, and despite counseling aided by available risk calculators, there are significant challenges with shared decision-making regarding life support and transition to end-of-life care. Improving predictive models can help providers and families navigate these unique challenges.

Objective: Machine learning methods have previously demonstrated added predictive value for determining intensive care unit outcomes, and their use allows consideration of a greater number of factors that potentially influence newborn outcomes, such as maternal characteristics. Machine learning–based models were analyzed for their ability to predict the survival of extremely preterm neonates at initial admission.

Methods: Maternal and newborn information was extracted from the health records of infants born between 23 and 29 weeks of gestation in the Medical Information Mart for Intensive Care III (MIMIC-III) critical care database. Applicable machine learning models predicting survival during the initial NICU admission were developed and compared. The same type of model was also examined using only features that would be available prepartum for the purpose of survival prediction prior to an anticipated preterm birth. Features most correlated with the predicted outcome were determined when possible for each model.

Results: Of included patients, 37 of 459 (8.1%) expired. The resulting random forest model showed higher predictive performance than the frequently used Score for Neonatal Acute Physiology With Perinatal Extension II (SNAPPE-II) NICU model when considering extremely preterm infants of very low birth weight. Several other machine learning models were found to have good performance but did not show a statistically significant difference from previously available models in this study. Feature importance varied by model, and those of greater importance included gestational age; birth weight; initial oxygenation level; elements of the APGAR (appearance, pulse, grimace, activity, and respiration) score; and amount of blood pressure support. Important prepartum features also included maternal age, steroid administration, and the presence of pregnancy complications.

Conclusions: Machine learning methods have the potential to provide robust prediction of survival in the context of extremely preterm births and allow for consideration of additional factors such as maternal clinical and socioeconomic information. Evaluation of larger, more diverse data sets may provide additional clarity on comparative performance.

(JMIR Med Inform 2024;12:e42271) doi:10.2196/42271

KEYWORDS

reproductive informatics; pregnancy complications; premature birth; neonatal mortality; machine learning; clinical decision support; preterm; pediatrics; intensive care unit outcome; health care outcome; survival prediction; maternal health; decision tree model; socioeconomic

Introduction

Preterm birth has long been a leading cause of infant mortality, with the lowest gestational age births associated with the highest rates of mortality [1]. In 2019, 59,506 infants were born at 31 weeks or less in the United States, and the infant mortality rate in this cohort was 18% [2]. When a patient is expected to deliver an extremely preterm infant, counseling on possible outcomes, methods of resuscitation, and anticipated course in the neonatal intensive care unit (NICU) ideally begins prior to birth. Many providers have used the National Institute of Child Health and Human Development (NICHD) risk calculator to initiate this discussion on the chances of infant mortality and severe morbidity after birth. The calculator is based on a logistic regression model using 5 prepartum factors (gestational age, estimated weight, sex, antenatal steroids, and multiple birth), derived from the preterm birth data of a network of US hospitals. With advances in NICU care and more knowledge about long-term outcomes, the calculator was updated in 2020 and maintains a similar performance (mean 0.744, SD 0.005) [3,4]. After initial resuscitation, several scoring systems are also available to predict mortality after a neonate arrives in the NICU [5-7]. However, they are less predictive with extremely low birth weight infants, as evidenced by the Score for Neonatal Acute Physiology With Perinatal Extension II (SNAPPE-II) survival model having a mean performance of 0.78 (SD 0.01) for infants weighing less than 1500 g at birth versus 0.91 (SD 0.01) overall. On review of several models, Clinical Risk Index for Babies (CRIB) had the highest performance in predicting very low birth weight neonate survival, with a mean of 0.88 (SD 0.02), although the CRIB and SNAPPE models were developed with data from geographically separate populations (Europe vs North America) [8].

Despite counseling supported by available risk calculators, decisions surrounding the continuation of life support and redirection to end-of-life care remain extremely difficult in the context of birth at the periviable preterm gestational ages because the postnatal course can be highly variable [9-11]. In addition, perceptions regarding the clinical situation can differ among providers and family members, and consideration of clinical and social context may be helpful [12,13].

Numerous machine learning models have been tested to improve the prediction of adult intensive care unit outcomes. The Medical Information Mart for Intensive Care III (MIMIC-III) database, which contains electronic health record (EHR) information of critical care patients at the Beth Israel Deaconess Medical Center from 2001 to 2012, has often been a source of data used in their development and testing [14-17]. Using the NICU data from MIMIC-III, this study builds and compares different types of machine learning algorithms that predict neonatal mortality and examines the value of incorporating features representing both structured and unstructured clinical elements for extremely preterm infants.

Methods

Ethical Considerations

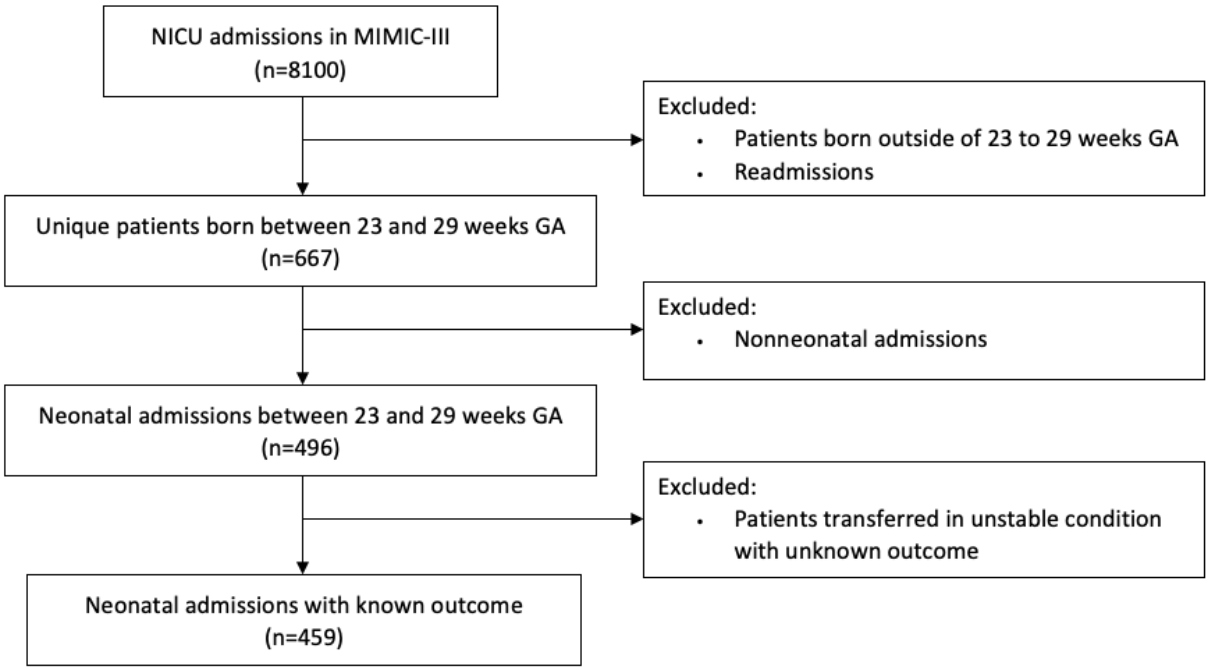
The institutional review board of the University at Buffalo determined the study (ID STUDY00003721) to be exempt as a secondary analysis of a publicly available data set. A data use agreement was obtained for the MIMIC-III database, which contains deidentified protected health information freely available for secondary analysis. The primary data collection for MIMIC-III was originally approved by the institutional review boards of Beth Israel Deaconess Medical Center and Massachusetts Institute of Technology with a waiver of individual patient consent, and no compensation was provided at that time.

Data Selection

Records of extremely preterm neonates admitted to the NICU in the MIMIC-III database were extracted using PostgreSQL (The PostgreSQL Global Development Group). A query was performed for admissions with *ICD-9* (*International Classification of Diseases, Ninth Revision*) codes corresponding to extremely preterm delivery less than 30 weeks as well as very low birth weight. From the resulting records, those of neonates born outside of 23 to 29 weeks were excluded, as well as duplicate records and readmissions. Some records corresponded to nonneonatal admissions, for example, where an infant had a prior history of preterm birth, and they were excluded. When the remaining records were reviewed, it was found that some neonates were transferred outside of the hospital for surgery and had an unknown outcome. These records were also excluded (Figure 1).

From the 459 neonatal admission records that were selected, the patients' demographics, vital signs, laboratory results, medications, procedures, and clinical text were queried from the database and reviewed. Of the available information, relevant elements were extracted based on factors found to be pertinent in previous scoring systems and expert knowledge. By manually curating the clinical text, including completed admission and discharge notes, we were able to incorporate features found only in unstructured form, including maternal clinical comorbidities and pregnancy complications. For this study, consideration of neonatal assessment and treatment was limited to data found initially at the time of NICU admission. The nonnumerical elements were encoded. Data that varied by clinical severity were encoded in that order, and the remaining categorical data underwent binary encoding. Median imputation was used to complete missing data.

Figure 1. Flowchart of selection criteria. GA: gestational age; MIMIC-III: Medical Information Mart for Intensive Care III; NICU: neonatal intensive care unit.



Ultimately, 83 features that could be used in machine learning algorithms were generated, of which approximately half represented maternal clinical and demographic information, with the remaining features representing infant findings at the time of admission (Multimedia Appendix 1).

Model Analysis

Several machine learning classification algorithms were implemented using Python 3.8 scikit-learn 1.2, and the resulting models were tested for their efficacy in predicting mortality. The same algorithms were also examined considering only prepartum features, assuming birth weight would be an estimated weight, to produce models that could be of assistance for clinicians counseling patients prior to an extremely preterm birth.

The performance of each model was endeavored to be optimized. To ensure that feature value range did not drive performance, standard scaling as well as min-max scaling were applied to quantitative features and used for models that were dependent upon distance calculations (eg, logistic regression, neural network, and support vector machine [SVM]). The final reported models used standard scaling due to improved performance over min-max scaling. Scaling was not performed for models invariant to monotonic transformations, such as random forest [18]. For the decision tree–based models, the hyperparameters of number of trees and maximum depth were adjusted. Number of trees began at 50 estimators and was increased by 50 until performance plateaued, which was at 250 trees with a maximum depth of 6 for the random forest method and 350 trees with a maximum depth of 5 for AdaBoost. The k value in the k-nearest neighbor algorithm was adjusted from the default value of 3 up to 20 (approximating the square root of the number of samples), and performance peaked at 4 in the final model. Because of the expected relatively small and imbalanced class sizes (8.1% in the minority class), a held-out test set was not used, and 10-fold stratified cross-validation with an 80:20 training and testing ratio was performed to ensure similar ratios across folds [19]. Mean performance metrics for F_1-score, area under the receiver operating characteristic (AUROC), and average precision are reported, as well as log loss and Brier score, where a smaller value is ideal when considering imbalanced classification.

Features most correlated with the predicted outcome were determined for the higher-performing methods. For the logistic regression model, coefficients most positively and negatively associated with mortality could be determined. For the remaining machine learning models, the most influential features were either directly queried using an available scikit-learn method or through the calculation of feature permutation importance.

Results

Of the included neonatal patients, 37 of 459 (8.1%) expired during the admission period after birth. The average length of stay for infants who survived after initial admission was 62.5 (SD 37.3) days. The average gestational age of the neonates at birth was 27 (SD 1.67) weeks, and 236 (51.4%) were male versus 223 (48.6%) female. Birth weights ranged from 365 to 2165 g, with the average birth weight being 1016 (SD 278) g, and 441 neonates were considered to have a very low birth weight (<1500 g). The average maternal age was 31.4 (SD 6.02) years. In terms of race and ethnicity, the majority of the included infants were in a category considered to be White (n=278, 60.1%), followed by Black (n=69, 15%), unknown (n=42, 9.2%), other (n=25, 5.4%), Hispanic (n=25, 5.4%), Asian (n=16, 3.5%), and Native American (n=4, 0.9%; Table 1).

Table 1. Demographics of patients whose records were included in the study.

	Total (N=459), n (%)	Survived (n=422, 91.9%), n (%)	Expired (n=37, 8.1%), n (%)
Gestational age (weeks)			
23	7 (1.5)	2 (28.6)	5 (71.4)
24	40 (8.7)	28 (70)	12 (30)
25	41 (8.9)	36 (87.8)	5 (12.2)
26	52 (11.3)	49 (94.2)	3 (5.8)
27	87 (19)	84 (96.6)	3 (3.4)
28	106 (23.1)	98 (92.5)	8 (7.5)
29	126 (27.5)	125 (99.2)	1 (0.8)
Sex			
Male	236 (51.4)	214 (90.7)	22 (9.3)
Female	223 (48.6)	208 (93.3)	15 (6.7)
Race			
Asian	16 (3.5)	15 (93.7)	1 (6.3)
Black	69 (15)	62 (89.9)	7 (10.1)
Hispanic	25 (5.4)	23 (92)	2 (8)
Native American	4 (0.9)	3 (75)	1 (25)
White	278 (60.1)	255 (91.7)	23 (8.3)
Other	25 (5.4)	23 (92)	2 (8)
Unknown	42 (9.2)	41 (97.6)	1 (2.4)
Insurance			
Private	343 (74.7)	311 (90.7)	32 (9.3)
Government	116 (25.3)	113 (97.4)	3 (2.6)
Uninsured	2 (0.4)	0 (0)	2 (100)
Family religion			
Catholic	100 (21.8)	91 (91)	9 (9)
Protestant	24 (5.2)	22 (91.7)	2 (8.3)
Jewish	16 (3.5)	15 (93.7)	1 (6.3)
Other	30 (6.5)	25 (83.3)	5 (16.7)
Unknown	289 (63)	269 (93.1)	20 (6.9)
Type of delivery			
Cesarean section	356 (77.6)	331 (93)	25 (7)
Vaginal delivery	103 (22.4)	91 (88.3)	12 (11.7)
Pregnancy type			
Singleton	247 (53.8)	230 (93.1)	17 (6.9)
Multiple	212 (46.2)	192 (90.6)	20 (9.4)
Antenatal steroids			
Received	369 (80.4)	347 (94)	22 (6)
Partially received	71 (15.5)	65 (91.5)	6 (8.5)
Not received	19 (4.1)	14 (73.7)	5 (26.3)

Logistic regression, Naïve Bayes, k-nearest neighbor, SVM, random forest, AdaBoost, and neural network classifiers were compared for efficacy in predicting mortality (Figure 2 and Table 2). Standard scaling transformation improved performance only for the logistic regression, SVM, and neural network methods. The random forest model had the highest predictive

performance when considering overall AUROC (mean 0.91, SD 0.07), F_1-score (0.67), and Brier score (0.06). The AdaBoost model had the next highest AUROC (mean 0.88, SD 0.10); however, the F_1-score (0.45) was low due to poor precision. On the other hand, the neural network model yielded the top F_1-score (0.67) and Brier score (0.05) despite having a lower AUROC (mean 0.84, SD 0.16). SVM was overall next best performing model (mean 0.86, SD 0.13; F_1-score 0.62; Brier score 0.06), followed by logistic regression (mean 0.82, SD 0.16; F_1-score 0.61; Brier score 0.08). The Naïve Bayes (mean 0.74, SD 0.22; F_1-score 0.40; Brier score 0.25) and k-nearest neighbor (mean 0.64, SD 0.13; F_1-score 0.34; Brier score 0.07) methods were the worst performing.

Figure 2. Receiver operating characteristic curves for the highest-performing models in Table 2. A: Logistic regression; B: SVM (support vector machine); C: Random forest; D: AdaBoost; E: Neural networks, F: Naïve Bayes; AUROC: area under the receiver operating characteristic; FP: false positive; TP: true positive.

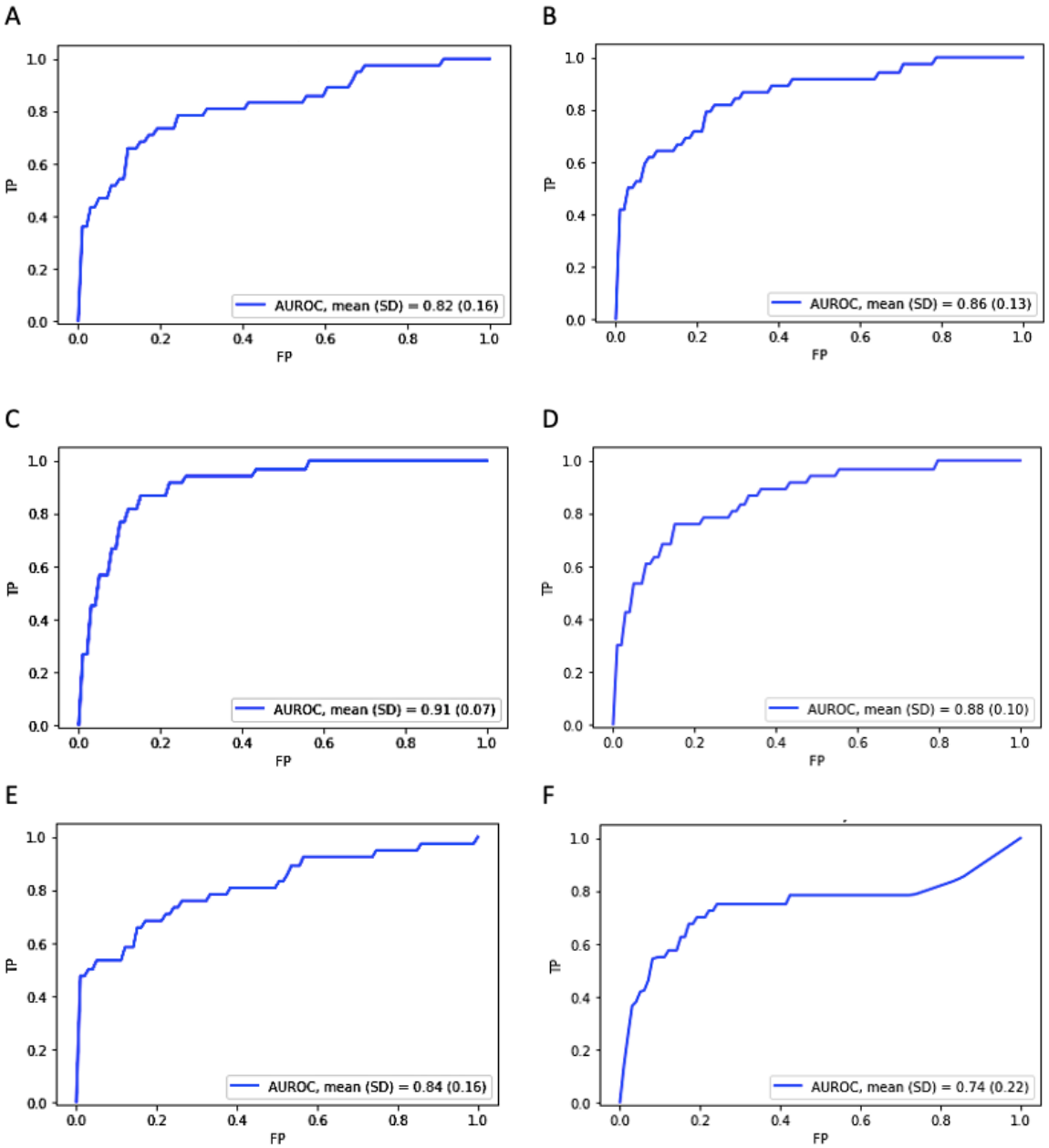


Table 2. AUROC[a], average precision, F1-score, log loss, and Brier scores for 10-fold stratified cross-validation predicting mortality using initial neonatal intensive care unit admission features (lower log loss and Brier scores are ideal when considering imbalanced classification).

Method	AUROC, mean (SD)	Precision, mean (SD)	F_1-score	Log loss score	Brier score
Logistic regression	0.82 (0.16)	0.55 (0.25)	0.61	0.35	0.08
SVM[b]	0.86 (0.13)	0.61 (0.24)	0.62	0.20	0.06
Random forest	0.91 (0.07)	0.61 (0.22)	0.67	0.19	0.06
AdaBoost	0.88 (0.10)	0.55 (0.25)	0.45	0.80	0.07
Neural network	0.84 (0.16)	0.65 (0.24)	0.67	0.30	0.05
Naïve Bayes	0.74 (0.22)	0.39 (0.17)	0.40	3.90	0.25
K-nearest neighbor	0.64 (0.13)	0.24 (0.16)	0.34	1.74	0.07

[a]AUROC: area under the receiver operating characteristic.

[b]SVM: support vector machine.

On post hoc chi-square analysis of the categorical variables, the factors that most influenced the outcome were insurance status, initial breathing assessment of the infant, and presence of a serious fetal anomaly (Table 3). When examining Pearson correlation of continuous variables, higher levels of ventilation and blood pressure support as well as higher arterial blood gas base deficit were properties mildly to moderately correlated with mortality. Larger gestational age, birth weight, and higher APGAR (appearance, pulse, grimace, activity, and respiration) scores at birth negatively correlated with mortality to a similar degree (Table 4).

Similar features were most strongly associated with outcome in the machine learning–based models, although they varied in importance (Table 5). For example, in the random forest model, gestational age, birth weight, and initial oxygen level were of higher importance, whereas in the neural network model, initial blood pressure support and activity level were the most influential features.

Evaluation of classifiers using only prepartum features, assuming birth weight as the estimated weight, also yielded the highest performance measures with the random forest method (Table 6). The random forest features that were consistently of highest importance included gestational age, weight, and maternal age (Table 7).

Table 3. Chi-square: categorical features significantly associated with outcome.

Feature	Description	Chi-square (*df*)
un_ins	Uninsured	22.8 (1)
breathing1	Initial breathing assessment	21.4 (2)
anomaly	Serious fetal anomaly	20.8 (1)
airway1	Initial type of airway or ventilation	17.1 (4)
religion_jehovahs	Religion Jehovah's Witness	11.4 (1)
twintwin	Twin-twin transfusion syndrome	9.4 (1)
uncertain	Uncertain pregnancy dating	7.9 (1)
religion_other	Religion other	4.9 (1)
gov_ins	Medicaid or Medicare insurance	4.7 (1)
muscle1	Muscle tone	4.6 (4)

Table 4. Pearson correlation: correlation of continuous features with mortality.

Feature	Description	Correlation
FiO2_1	Initial amount of oxygen ventilation	0.28
BD1	Initial arterial blood gas base deficit	0.23
dopa1	Initial IV[a] dopamine rate	0.20
temp1	Initial temperature	0.14
pCO2_1	Initial arterial blood gas pCO_2[b]	0.13
G	Maternal gravidity	0.07
P	Maternal parity	0.06
PRBC1	Initial IV blood transfusion amount	0.06
maternal_age	Maternal age	0.05
gluc1	Initial glucose	0.03
bands1	Initial bands	0.03
multiple	Number of fetuses at delivery	0.02
pO2_1	Initial arterial blood gas pO_2[c]	–0.01[d]
wbc1	Initial white blood cells	–0.02
BPmean1	Initial mean blood pressure	–0.04
monos1	Initial monocytes	–0.05
HR1	Initial heart rate	–0.05
hct1	Initial hematocrit	–0.07
neuts1	Initial neutrophil count	–0.07
SaO2_1	Initial oxygen saturation	–0.20
birth_wt	Birth weight	–0.22
GA	Gestational age at birth	–0.32
apgar1	One-minute APGAR[e] score	–0.32
apgar5	Five-minute APGAR score	–0.35

[a]IV: intravenous.

[b]pCO_2: partial pressure of carbon dioxide

[c]pO_2: partial pressure of oxygen.

[d]Negative correlations with mortality imply a correlation with survival.

[e]APGAR: appearance, pulse, grimace, activity, and respiration.

Table 5. Features of highest importance in various models, listed in order of importance. Positive and negative associations with mortality can be calculated only in logistic regression models. For the tree-based random forest and AdaBoost algorithms, an impurity-based method was used to determine overall feature importance. For the remaining algorithms, importance was found via feature permutation[a].

Logistic regression: positively associated with mortality	Logistic regression: negatively associated with mortality	Random forest	AdaBoost	SVM[b]	Neural network
race_hispanic	GA	GA	neuts1	activity1	dopa1
color1	race_unk	birth_wt	hct1	GA	activity1
anomaly	apgar1	SaO2_1	SaO2_1	HTN	multiple
race_asian	gov_ins	BD1	wbc1	anomaly	uncertain
un_ins	activity1	apgar1	apgar1	breathL1	race_unk
dopa1	monos1	gluc1	monos1	breathR1	twintwin
abdomen1	breathL1	dopa1	temp1	twintwin	anomaly
pvt_ins	PRBC1	apgar5	HR1	birth_wt	muscle1
multiple	infert	FiO2_1	FiO2_1	antfont1	wbc1
FiO2_1	dm	neuts	bands1	caprefill1	abdomen1

[a]The descriptions of variable names are present in Multimedia Appendix 1.

[b]SVM: support vector machine.

Table 6. AUROC[a], average precision, F1-score, log loss, and Brier scores for 10-fold stratified cross-validation predicting mortality when only prepartum features are available (lower log loss and Brier scores are ideal when considering imbalanced classification).

Method	AUROC, mean (SD)	Precision, mean (SD)	F_1-score	Log loss score	Brier score
Logistic regression	0.77 (0.14)	0.41 (0.18)	0.51	0.29	0.07
SVM[b]	0.76 (0.10)	0.37 (0.15)	0.46	0.25	0.07
Random forest	0.80 (0.14)	0.54 (0.27)	0.59	0.22	0.06
AdaBoost	0.75 (0.17)	0.44 (0.29)	0.54	0.27	0.07
Neural network	0.76 (0.11)	0.44 (0.18)	0.53	0.31	0.07
Naïve Bayes	0.68 (0.21)	0.30 (0.11)	0.19	6.09	0.59
K-nearest neighbor	0.62 (0.12)	0.20 (0.12)	0.30	1.77	0.09

[a]AUROC: area under the receiver operating characteristic.

[b]SVM: support vector machine.

Table 7. Prepartum features of highest importance in various models, listed in order of importance[a].

Logistic regression: positively associated with mortality	Logistic regression: negatively associated with mortality	Random forest	AdaBoost	SVM[b]	Neural network
maternal_age	GA	GA	birth_wt	GA	un_ins
anomaly	race_unk	birth_wt	maternal_age	steroids	steroids
un_ins	dm	maternal_age	GA	P	HTN
asthma	depression	anomaly	G	infert	GA
religion_jehovahs	PTL	G	multiple	G	anomaly
pvt_ins	steroids	P	religion_unk	uncertain	twintwin
race_hispanic	gov_ins	un_ins	steroids	birth_wt	race_unk
twintwin	HTN	steroids	sex	sex	sex
uncertain	infert	uncertain	anomaly	multiple	P
multiple	P	twintwin	P	anomaly	SVD

[a]The descriptions of variable names are present in Multimedia Appendix 1.

[b]SVM: support vector machine.

Several of the important features found in the top-performing models were among those manually curated in unstructured form, including the presence of maternal hypertensive disease and diabetes, uncertain pregnancy dating (uncertain), fetal anomaly (anomaly), and twin-twin transfusion syndrome.

Discussion

Principal Findings

There is a potential for existing risk calculators to be outperformed by tree-based machine learning algorithms, as indicated by the higher performance of our random forest model versus SNAPPE-II in the context of extremely premature or very low birth weight infants (in fact AUROC increased to mean 0.92, SD 0.05 when only the neonates <1500 g were considered in the random forest model to directly compare to SNAPPE-II). Performance difference compared with CRIB is inconclusive, however. In terms of estimating neonatal mortality prior to preterm birth, although the point estimates of several of the machine learning algorithms using additional features extracted from the EHR were higher than that of the NICHD calculator, overlapping CIs preclude any conclusion about significant differences in performance.

Comparison to Prior Work

Examination of prior work further points to the importance of using data available from the EHR, including unstructured health data. For example, the relatively high-performing CRIB score includes the presence of fetal malformation as a variable. Saria et al [20] incorporated signal processing of short-term time series data from neonatal vital sign sensors to produce a model classifying infants at high risk for severe morbidity or mortality. To maintain accuracy over time, Meadow et al [11] proposed a longitudinal NICU survival model combining adverse events, imaging report information, and caretaker intuition. Hamilton et al [21] more recently applied tree-based machine learning in the context of preterm birth to determine clusters of pregnancy characteristics that were at the highest risk for severe neonatal morbidity or mortality.

Strengths and Limitations

This study is limited by a small data set with data from a single institution, which in turn limits the ability to establish statistical significance in performance differences and the variety of machine learning methods that can be examined. Because of the retrospective nature of the study, there is less control over the format of the data and the amount of missing data. Although a single-institution data set is usually considered a limitation, Rysavy et al [4] emphasized that extremely preterm neonatal outcomes are significantly influenced by the hospital of birth and suggested maintaining ongoing and updated prediction models from outcomes within hospital systems. Using machine learning would be ideal for this task, allowing for consideration of a number of features retrievable from the EHR with a high tolerance for missing or outlier data as the volume of data increases. Tree-based machine learning algorithms may be additionally advantageous due to their ability to iteratively combine numerous weakly predictive features into stronger predictors.

Knowledge of the most influential features, which was possible to visualize in the majority of the presented models, provides transparency. Understanding which factors contribute most to the prediction of outcomes in a model can help clinical providers derive greater intuition regarding how applicable the model is to a particular patient.

The inclusion of maternal information and pregnancy characteristics found in unstructured form in the MIMIC-III database allowed for consideration of factors beyond the numerical neonatal data. Some of these additional variables, such as the presence of fetal anomalies or twin-twin transfusion syndrome, were found to be of high importance in several top-performing models, especially in those used in the prepartum period prior to an anticipated extremely preterm delivery. This illustrates that machine learning–based models could potentially be helpful for continuity of care, starting in the prepartum timeframe with ongoing predictive ability after birth. Maternal demographic information had an influence on mortality prediction in some of the higher-performing models but not others. Although demographic data can provide additional knowledge of social context, unintended bias can also be introduced into the resulting model [22].

Future Directions

Future work anticipates further evaluation of these methods on larger, more diverse data sets to determine if there is a significant and reproducible performance advantage. Expanding the study to include additional data would also allow the evaluation of more powerful machine learning methods such as deep learning methods. Eventually, the maintenance of a more representative and up-to-date cohort for training could potentially be accomplished via collaborative or federated learning techniques across institutions [22,23]. To address the possibility of algorithmic bias, further work could include a comparison of prediction results using models with and without protected demographic features and a calculation of the level of discrimination that could result. Assessment of more data from underrepresented groups may also aid in producing increasingly accurate and less discriminatory models [24,25].

In this study, unstructured information was manually extracted from admission and discharge notes in the MIMIC-III database and allowed for consideration of additional relevant features in our models. This suggests that the use of natural language processing to better understand clinical context may further improve the prediction of outcomes of extremely preterm births. As automated natural language processing of clinical notes becomes more mature and prevalent, the use of these features gleaned from unstructured EHR data will be increasingly applicable [26].

Additional potential future directions include integrating with or adding functionalities found in other intensive care unit models, such as time series modeling, and predicting outcomes other than mortality, such as the development of comorbidities, discharge location, length of stay, and likelihood of readmission.

Conclusions

This study examined machine learning models produced from the MIMIC-III NICU data set and their predictive ability in the

clinically challenging situation of extremely preterm birth. The tree-based random forest model was found to have higher performance than the SNAPPE-II model when predicting the survival of extremely preterm infants of very low birth weight. Several other models, including those using only features that would be known prepartum, also appeared to have good predictive performance but failed to show a statistically significant difference from prior models. Features of highest importance in these models were explored and included traditional variables, such as gestational age and birth weight, but also information that may be found in unstructured form in the EHR. Evaluation of these and even more advanced machine learning methods on larger data sets may offer further clarity about performance differences, and natural language processing techniques would allow for greater use of unstructured clinical information.

Acknowledgments

This work was supported by a National Institutes of Health National Library of Medicine training grant (T15 LM012495-02). Generative artificial intelligence was not used in any portion of the paper writing.

Data Availability

The data sets analyzed in this study are available in the Medical Information Mart for Intensive Care III (MIMIC-III) Clinical Database [14].

Authors' Contributions

AL and PLE conceptualized study methodology. AL and SM participated in data curation, statistical analysis, and writing. All authors reviewed the final paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Categorical and continuous features.
[DOCX File , 21 KB - medinform_v12i1e42271_app1.docx]

Multimedia Appendix 2
TRIPOD checklist for model development.
[PDF File (Adobe PDF File), 825 KB - medinform_v12i1e42271_app2.pdf]

References

1. Stoll BJ, Hansen NI, Bell EF, Walsh MC, Carlo WA, Shankaran S, et al. Trends in care practices, morbidity, and mortality of extremely preterm neonates, 1993-2012. JAMA 2015;314(10):1039-1051 [FREE Full text] [doi: 10.1001/jama.2015.10244] [Medline: 26348753]
2. Ely D, Driscoll A. Infant mortality in the United States, 2019: data from the period linked birth/infant death file. Natl Vital Stat Rep 2021;70(14):1-17 [FREE Full text] [doi: 10.15620/cdc:111053]
3. Dance A. Survival of the littlest: the long-term impacts of being born extremely early. Nature 2020;582(7810):20-23. [doi: 10.1038/d41586-020-01517-z] [Medline: 32488165]
4. Rysavy MA, Horbar JD, Bell EF, Li L, Greenberg LT, Tyson JE, et al. Assessment of an updated neonatal research network extremely preterm birth outcome model in the Vermont Oxford Network. JAMA Pediatr 2020;174(5):e196294 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.6294] [Medline: 32119065]
5. Richardson DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: simplified newborn illness severity and mortality risk scores. J Pediatr 2001;138(1):92-100. [doi: 10.1067/mpd.2001.109608] [Medline: 11148519]
6. Parry G, Tucker J, Tarnow-Mordi W, UK Neonatal Staffing Study Collaborative Group. CRIB II: an update of the clinical risk index for babies score. Lancet 2003;361(9371):1789-1791. [doi: 10.1016/S0140-6736(03)13397-1] [Medline: 12781540]
7. Groenendaal F, de Vos MC, Derks JB, Mulder EJH. Improved SNAPPE-II and CRIB II scores over a 15-year period. J Perinatol 2017;37(5):547-551. [doi: 10.1038/jp.2016.276] [Medline: 28125092]
8. McLeod JS, Menon A, Matusko N, Weiner GM, Gadepalli SK, Barks J, et al. Comparing mortality risk models in VLBW and preterm infants: systematic review and meta-analysis. J Perinatol 2020;40(5):695-703. [doi: 10.1038/s41372-020-0650-0] [Medline: 32203174]
9. Andrews B, Myers P, Lagatta J, Meadow W. A comparison of prenatal and postnatal models to predict outcomes at the border of viability. J Pediatr 2016;173:96-100. [doi: 10.1016/j.jpeds.2016.02.042] [Medline: 26995702]

10. Dupont-Thibodeau A, Barrington KJ, Farlow B, Janvier A. End-of-life decisions for extremely low-gestational-age infants: why simple rules for complicated decisions should be avoided. Semin Perinatol 2014;38(1):31-37. [doi: 10.1053/j.semperi.2013.07.006] [Medline: 24468567]
11. Meadow W, Lagatta J, Andrews B, Lantos J. The mathematics of morality for neonatal resuscitation. Clin Perinatol 2012;39(4):941-956 [FREE Full text] [doi: 10.1016/j.clp.2012.09.013] [Medline: 23164189]
12. Hellmann J, Knighton R, Lee SK, Shah PS, Canadian Neonatal Network End of Life Study Group. Neonatal deaths: prospective exploration of the causes and process of end-of-life decisions. Arch Dis Child Fetal Neonatal Ed 2016;101(2):F102-F107. [doi: 10.1136/archdischild-2015-308425] [Medline: 26253166]
13. Steurer MA, Anderson J, Baer RJ, Oltman S, Franck LS, Kuppermann M, et al. Dynamic outcome prediction in a socio-demographically diverse population-based cohort of extremely preterm neonates. J Perinatol 2017;37(6):709-715. [doi: 10.1038/jp.2017.9] [Medline: 28206998]
14. 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]
15. Tang F, Xiao C, Wang F, Zhou J. Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA Open 2018;1(1):87-98 [FREE Full text] [doi: 10.1093/jamiaopen/ooy011] [Medline: 31984321]
16. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 2018;83:112-134 [FREE Full text] [doi: 10.1016/j.jbi.2018.04.007] [Medline: 29879470]
17. Caicedo-Torres W, Gutierrez J. ISeeU: visually interpretable deep learning for mortality prediction inside the ICU. J Biomed Inform 2019;98:103269 [FREE Full text] [doi: 10.1016/j.jbi.2019.103269] [Medline: 31430550]
18. Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. Technologies 2021;9(3):52 [FREE Full text] [doi: 10.3390/technologies9030052]
19. 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 [FREE Full text]
20. Saria S, Rajani AK, Gould J, Koller D, Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. Sci Transl Med 2010;2(48):48ra65 [FREE Full text] [doi: 10.1126/scitranslmed.3001304] [Medline: 20826840]
21. Hamilton EF, Dyachenko A, Ciampi A, Maurel K, Warrick PA, Garite TJ. Estimating risk of severe neonatal morbidity in preterm births under 32 weeks of gestation. J Matern Fetal Neonatal Med 2020;33(1):73-80. [doi: 10.1080/14767058.2018.1487395] [Medline: 29886760]
22. Crowson MG, Moukheiber D, Arévalo AR, Lam BD, Mantena S, Rana A, et al. A systematic review of federated learning applications for biomedical data. PLOS Digit Health 2022;1(5):e0000033 [FREE Full text] [doi: 10.1371/journal.pdig.0000033] [Medline: 36812504]
23. Nguyen TV, Dakka MA, Diakiw SM, VerMilyea MD, Perugini M, Hall JMM, et al. A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data. Sci Rep 2022;12(1):8888 [FREE Full text] [doi: 10.1038/s41598-022-12833-x] [Medline: 35614106]
24. Chen IY, Johansson FD, Sontag D. Why is my classifier discriminatory? 2018 Presented at: Proceedings of the 32nd International Conference on Neural Information Processing Systems; December 3-8, 2018; Montréal, Canada p. 3543-3554. [doi: 10.5555/3327144.3327272]
25. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2021;4:123-144 [FREE Full text] [doi: 10.1146/annurev-biodatasci-092820-114757] [Medline: 34396058]
26. Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. J Am Med Inform Assoc 2022;29(7):1292-1302 [FREE Full text] [doi: 10.1093/jamia/ocac058] [Medline: 35475536]

Abbreviations

APGAR: appearance, pulse, grimace, activity, and respiration
AUROC: area under the receiver operating characteristic
CRIB: Clinical Risk Index for Babies
EHR: electronic health record
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
MIMIC-III: Medical Information Mart for Intensive Care III
NICHD: National Institute of Child Health and Human Development
NICU: neonatal intensive care unit
SNAPPE-II: Score for Neonatal Acute Physiology With Perinatal Extension II
SVM: support vector machine

Edited by C Lovis; submitted 30.08.22; peer-reviewed by M Casal-Guisande, F Meza; comments to author 17.11.22; revised version received 02.02.23; accepted 28.12.23; published 14.02.24.

Please cite as:
Li A, Mullin S, Elkin PL
Improving Prediction of Survival for Extremely Premature Infants Born at 23 to 29 Weeks Gestational Age in the Neonatal Intensive Care Unit: Development and Evaluation of Machine Learning Models
JMIR Med Inform 2024;12:e42271
URL: https://medinform.jmir.org/2024/1/e42271
doi:10.2196/42271
PMID:38354033

Corrigenda and Addenda

Correction: A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study

Zhixiang Zhao[1,2*], MD, PhD; Che-Ming Wu[3*], MEng; Shuping Zhang[1,2], MD, PhD; Fanping He[1,2], MD, PhD; Fangfen Liu[1,2], MD, PhD; Ben Wang[1,2], MD, PhD; Yingxue Huang[1,2], MD, PhD; Wei Shi[1,2], MD, PhD; Dan Jian[1,2], MD, PhD; Hongfu Xie[1,2], MD, PhD; Chao-Yuan Yeh[3*], MD; Ji Li[1,2,4,5*], MD, PhD

[1]Department of Dermatology, Xiangya Hospital of Central South University, Changsha, China

[2]Hunan Key Laboratory of Aging Biology, Xiangya Hospital of Central South University, Changsha, China

[3]aetherAI, Co Ltd, Taipei, Taiwan, China

[4]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital of Central South University, Changsha, China

[5]Key Laboratory of Organ Injury, Aging and Regenerative Medicine of Hunan Province, Changsha, China

[*]these authors contributed equally

Corresponding Author:
Ji Li, MD, PhD
Department of Dermatology
Xiangya Hospital of Central South University
87 Xiangya Rd.
Changsha, 410008
China
Phone: 86 073189753406
Email: liji_xy@csu.edu.cn

Related Article:

Correction of: https://medinform.jmir.org/2021/3/e23415/

(JMIR Med Inform 2024;12:e57654) doi:10.2196/57654

In “A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study” (JMIR Med Inform 2021;9(3):e23415) the authors made one addition.

An “Acknowledgments” section has been added that reads as follows:

> *This work was supported by The Educational Science and Planning Project of Hunan Province (XTK20BGD008).*

The correction will appear in the online version of the paper on the JMIR Publications website on March 8, 2024, 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 26.02.24; this is a non–peer-reviewed article;accepted 27.02.24; published 08.03.24.

Please cite as:
Zhao Z, Wu CM, Zhang S, He F, Liu F, Wang B, Huang Y, Shi W, Jian D, Xie H, Yeh CY, Li J
Correction: A Novel Convolutional Neural Network for the Diagnosis and Classification of Rosacea: Usability Study
JMIR Med Inform 2024;12:e57654
URL: https://medinform.jmir.org/2024/1/e57654
doi:10.2196/57654
PMID:38457810

Corrigenda and Addenda

Correction: 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:

Correction of: https://medinform.jmir.org/2023/1/e44892

(JMIR Med Inform 2024;12:e62757) doi:10.2196/62757

In "A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification" (JMIR Med Inform 2023;11:e44892), the authors made one addition.

An Acknowledgments section was added to the paper, as follows:

> *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).*

The correction will appear in the online version of the paper on the JMIR Publications website on June 5, 2024, 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 30.05.24; this is a non–peer-reviewed article;accepted 03.06.24; published 05.06.24.

Please cite as:
Zhang Y, Li X, Liu Y, Li A, Yang X, Tang X
Correction: A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification
JMIR Med Inform 2024;12:e62757
URL: https://medinform.jmir.org/2024/1/e62757
doi:10.2196/62757
PMID:38838306

Correction: 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:

Correction of: https://medinform.jmir.org/2023/1/e53112

(JMIR Med Inform 2024;12:e56050) doi:10.2196/56050

In "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) the author noted one error.

In the section titled "The Status Quo," the following sentence appears:

> *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 people in England alone.*

This has been changed to read as follows:

> *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.*

The correction will appear in the online version of the paper on the JMIR Publications website on January 12, 2024 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 03.01.24; this is a non–peer-reviewed article;accepted 03.01.24; published 12.01.24.

Please cite as:
Morris JS
Correction: A Call to Reconsider a Nationwide Electronic Health Record System: Correcting the Failures of the National Program for IT
JMIR Med Inform 2024;12:e56050
URL: https://medinform.jmir.org/2024/1/e56050
doi:10.2196/56050

Development of a Trusted Third Party at a Large University Hospital: Design and Implementation Study

Eric Wündisch[1], MSc; Peter Hufnagl[2], Prof Dr; Peter Brunecker[3], Dr rer medic; Sophie Meier zu Ummeln[1], CEng; Sarah Träger[1], MA; Marcus Kopp[1], BSc; Fabian Prasser[4,*], Prof Dr; Joachim Weber[1,5,6,*], Dr med

1

2

3

4

5

6

[*]these authors contributed equally

Corresponding Author:
Eric Wündisch, MSc

Abstract

Background: Pseudonymization has become a best practice to securely manage the identities of patients and study participants in medical research projects and data sharing initiatives. This method offers the advantage of not requiring the direct identification of data to support various research processes while still allowing for advanced processing activities, such as data linkage. Often, pseudonymization and related functionalities are bundled in specific technical and organization units known as trusted third parties (TTPs). However, pseudonymization can significantly increase the complexity of data management and research workflows, necessitating adequate tool support. Common tasks of TTPs include supporting the secure registration and pseudonymization of patient and sample identities as well as managing consent.

Objective: Despite the challenges involved, little has been published about successful architectures and functional tools for implementing TTPs in large university hospitals. The aim of this paper is to fill this research gap by describing the software architecture and tool set developed and deployed as part of a TTP established at Charité – Universitätsmedizin Berlin.

Methods: The infrastructure for the TTP was designed to provide a modular structure while keeping maintenance requirements low. Basic functionalities were realized with the free MOSAIC tools. However, supporting common study processes requires implementing workflows that span different basic services, such as patient registration, followed by pseudonym generation and concluded by consent collection. To achieve this, an integration layer was developed to provide a unified Representational state transfer (REST) application programming interface (API) as a basis for more complex workflows. Based on this API, a unified graphical user interface was also implemented, providing an integrated view of information objects and workflows supported by the TTP. The API was implemented using Java and Spring Boot, while the graphical user interface was implemented in PHP and Laravel. Both services use a shared Keycloak instance as a unified management system for roles and rights.

Results: By the end of 2022, the TTP has already supported more than 10 research projects since its launch in December 2019. Within these projects, more than 3000 identities were stored, more than 30,000 pseudonyms were generated, and more than 1500 consent forms were submitted. In total, more than 150 people regularly work with the software platform. By implementing the integration layer and the unified user interface, together with comprehensive roles and rights management, the effort for operating the TTP could be significantly reduced, as personnel of the supported research projects can use many functionalities independently.

Conclusions: With the architecture and components described, we created a user-friendly and compliant environment for supporting research projects. We believe that the insights into the design and implementation of our TTP can help other institutions to efficiently and effectively set up corresponding structures.

(JMIR Med Inform 2024;12:e53075) doi:10.2196/53075

KEYWORDS

pseudonymisation; architecture; scalability; trusted third party; application; security; consent; identifying data; infrastructure; modular; software; implementation; user interface; health platform; data management; data privacy; health record; electronic health record; EHR; pseudonymization

Introduction

Background

Medical research relies on the effective collection, management, and analysis of biomedical data [1]. However, the complexity of associated data flows is increasing constantly due to the rising importance of data-driven approaches from the areas of data science and artificial intelligence [2,3]. These typically require data to be reused and shared to generate the necessary large data sets, for example in neuroscience [4]. At the same time, relevant data are often highly sensitive and require protection against unauthorized use and disclosure [5]. In alignment with this need, various laws, regulations, guidelines, and best practices suggest pseudonymization as a central data protection mechanism, especially in biomedical research [6]. Pseudonymization refers to a process in which data that directly identifies individuals (henceforth denoted as identifying data), such as names and addresses, are stored separately from data and biosamples needed for scientific analyses, and research assets are identified using protected identifiers, known as pseudonyms [7]. This protects the identity of patients or study participants while still allowing the implementation of complex research workflows, for example, data linkage. It is frequently suggested to bundle pseudonymization with other functionalities relevant to data protection and compliance, such as consent management, and that those should be carried out by particularly trusted units, knwon as trusted third parties (TTPs). One example of a concept recommending TTPs is the Guideline for Data Protection in Medical Research Projects by Technology, Methods, and Infrastructure for Networked Medical Research (TMF), the German umbrella organization for networked medical research [8].

Although the general functionalities required by medical research projects may be similar, the way they are combined into workflows often differs significantly. The reason is that due to varying study schedules and (data) modalities, studies often have different requirements concerning the necessary number and types of pseudonyms as well as the research assets that have to be registered. The timing of consent collection can also vary, for example, if reconsenting is required. Another factor that can contribute to heterogeneity is the need for integration of or linkage with data from external systems or institutions. As a result, studies often develop study- or project-specific solutions to fulfill specific registration, pseudonymization, linkage, and consenting requirements [9]. Some open tools, such as Enterprise Identifier Cross-Referencing (E-PIX) [10], Generic Pseudonym Administration Service (gPAS) [11], Generic Informed Consent Service (gICS) [12], or Mainzelliste [13], have been developed and are in widespread use; however, they are usually not integrated with each other, making the implementation of more complex workflows involving different TTP operations challenging and potentially lead to systematic limitations (explained further in the *Discussion* section). Although research exists on the components mentioned above, the literature lacks insights into the design of more comprehensive architectures that support complex research workflows that are actually in production use [14,15].

Objectives

This paper presents the design of a comprehensive architecture for a TTP that aims to support a wide range of different research projects and studies using a unified system. As a first step, we present requirements elicited for this structure and then describe the implementation of a corresponding solution that reuses existing open components. These components are extended with a common application programming interface (API) and a common graphical user interface (GUI). We then present insights into our experiences with piloting this structure and describe our plans for future developments.

Methods

Requirements

TTPs typically offer a range of core functionalities based on their role in supporting research projects and clinical studies with data protection services. Three key functionalities provided are as follows: (1) identity management, through which patients and study participants are registered and their identities are managed across different systems using record linkage; (2) pseudonym management, which provides and manages pseudonyms for different research contexts and is thus critical for data protection compliance; and (3) consent management, to obtain and manage patient and participant consent for various research activities. Further components are usually included to make these core functionalities accessible. An API is necessary for the systematic retrieval of information, the implementation of complex workflows, and integration with further health care and research systems. Moreover, a well-designed GUI is necessary to enable TTP staff and study personnel to perform common tasks efficiently. An audit trail is required to ensure transparency and traceability. Furthermore, data import and export functions are necessary for transferring data from legacy systems and archiving in study-specific contexts. Finally, platform independence is an important nonfunctional requirement to support wide adoption.

A common set of tools providing these core functionalities and features (Table 1) are E-PIX [10], gPAS [11], and gICS [12], which are provided as free web-based software by the MOSAIC project from the University of Greifswald (explained in the following section). They are successfully used in a range of research projects and infrastructures [16]. Table 1 illustrates which of the above-mentioned core requirements are fulfilled by which of the MOSAIC tools.

Table . Core functional requirements and MOSAIC tools that fulfill them.

Core functional requirements		Tools		
		E-PIX[a]	gPAS[b]	gICS[c]
Basic services				
	Identity management	✓	—[d]	—
	Pseudonym management	—	✓	—
	Consent management	—	—	✓
Additional features				
	Application programming interface	✓	✓	✓
	Graphical user interface	✓	✓	✓
	Audit trail	✓	—	✓
	Data import and export	✓	✓	✓

[a]E-PIX: Enterprise Identifier Cross-Referencing.

[b]gPAS: Generic Pseudonym Administration Service.

[c]gICS: Generic Informed Consent Service.

[d]Not applicable.

Although the MOSAIC tools provide the basic functionalities needed, we elicited additional requirements from our extensive experience with supporting research projects. An overview is provided in Table 2. A detailed discussion is available in the section *Comparison With Prior Work*.

Table . Additional functional requirements and core services for which they are relevant.

Additional functional requirements		Identity management	Pseudonym management	Consent management
Programmatic interfaces and workflows				
	Modern REST[a] application programming interface	✓	✓	✓
	Information exchange with other systems (eg, for ingesting consents documented in the EHR[b] system)	✓	✓	✓
	Cross-system workflows (eg, creation of a primary identifier, combined with the creation of all necessary pseudonyms based on the domain tree and preparation of a consent document)	✓	✓	✓
User interfaces and services				
	Integrated user interface across all services	✓	✓	✓
	Common authentication and authorization framework with single-sign-on and associated rights and roles with the ability to connect to institutional directory services	✓	✓	✓
	Sending status messages to users in case of relevant events (eg, when a new patient has been registered)	✓	✓	✓
Specific features				
	Visualization of pseudonyms as QR codes	—[c]	✓	—
	Automated versioning when storing consent updates	—	—	✓
	Kiosk mode for consent documentation	—	—	✓

[a]REST: representational state transfer.

[b]EHR: electronic health record.

[c]Not applicable.

Programmatic Interfaces and Workflows

Representational state transfer (REST) services have become a de facto standard for modern applications over the last couple of years, as they are stateless, lean, and based on open web standards. Hence, we considered a REST API to be an important requirement for all 3 areas—identity management, pseudonym management, and consent management. Together with other common technologies, such asJavaScript Object Notation, this makes the services offered by the TTP accessible to other systems and processes. It also fosters effective information exchange with other systems, for example, to automatically generate primary identifiers and pseudonyms in case a patient is registered in the electronic health record (EHR) system. Moreover, a common API across all services also enables cross-service workflows, which we consider particularly important. An example of this is the automatic creation of pseudonyms linked to the primary identifier when registering a patient or study participant.

User Interfaces and Services

We considered an integrated user interface (UI) together with a shared authentication and authorization mechanism to be central for our TTP infrastructure. Important functionalities that the UI needs to support include depseudonymization, patient and participant registration, consent management and configuration, as well as administration. A tighter integration of the different components also facilitates sending status messages to users in case actions are required on their side.

Specific Features

We further identified requirements in regard to specific management functionalities. For example, representing pseudonyms as QR codes is important for seamless workflows

across different media; this includes printing the codes on accompanying documents or biospecimen tubes and then reading them using QR code readers. This is particularly important for biospecimen management. Moreover, we identified a need for versioning of managed consent documents. In the event of updates to consents, for example, due to wrong information on the consent form, versioning of the various consents in the system is important for traceability. This also requires the system to be able to assign consents or withdrawals to other participants (eg, if a wrong identifier has been used when originally collecting the form). In addition, a kiosk mode that locks the user into the application is needed for the secure collection of consents from patients using tablets.

Nonfunctional Requirements

The most important nonfunctional requirements are as follows: (1) scalability, particularly when executing cross-service operations, and (2) documentation of administration functions.

Building Blocks

In this section, we will describe basic building blocks of the developed application stack.

MOSAIC Tools

As mentioned previously, the application has been developed around the MOSAIC tools [17] as core components. Although these tools do not fulfill all our requirements, they provide a solid basis for implementing the core functionalities. The MOSAIC tools have been positively evaluated by the data protection authority of Mecklenburg-Vorpommern in Germany [18] and have been successfully used in several research projects, for example, the BeLOVE (Berlin Longterm Observation of Vascular Events) [19,20] and NAKO (German National Cohort) studies [21].

The MOSAIC suite consists of 3 tools [22]: E-PIX provides a master patient index following the Integrating the Healthcare Enterprise (IHE) profiles, Patient Identifier Cross-Reference (PIX), and Patient Demographics Query [23,24]; gPAS provides associated pseudonymization functionalities; and gICS supports integrated consent management. More specifically, E-PIX enables the central management of directly identifying master data and supports probabilistic record linkage. The resolution of potential matches between identifying data is supported through the UI. gPAS supports the generation and management of pseudonyms on top of the identities managed by E-PIX using different pseudonym domains that can refer to different systems, locations, or contexts. Finally, gICS supports digitally managing informed consent and supports different consent templates and associated use policies.

Following our requirements, we implemented an authentication and authorization model as well as programmatic interfaces and graphical UIs around E-PIX, gPAS, and gICS to enable integrated workflows across all 3 tools and to improve their interfaces.

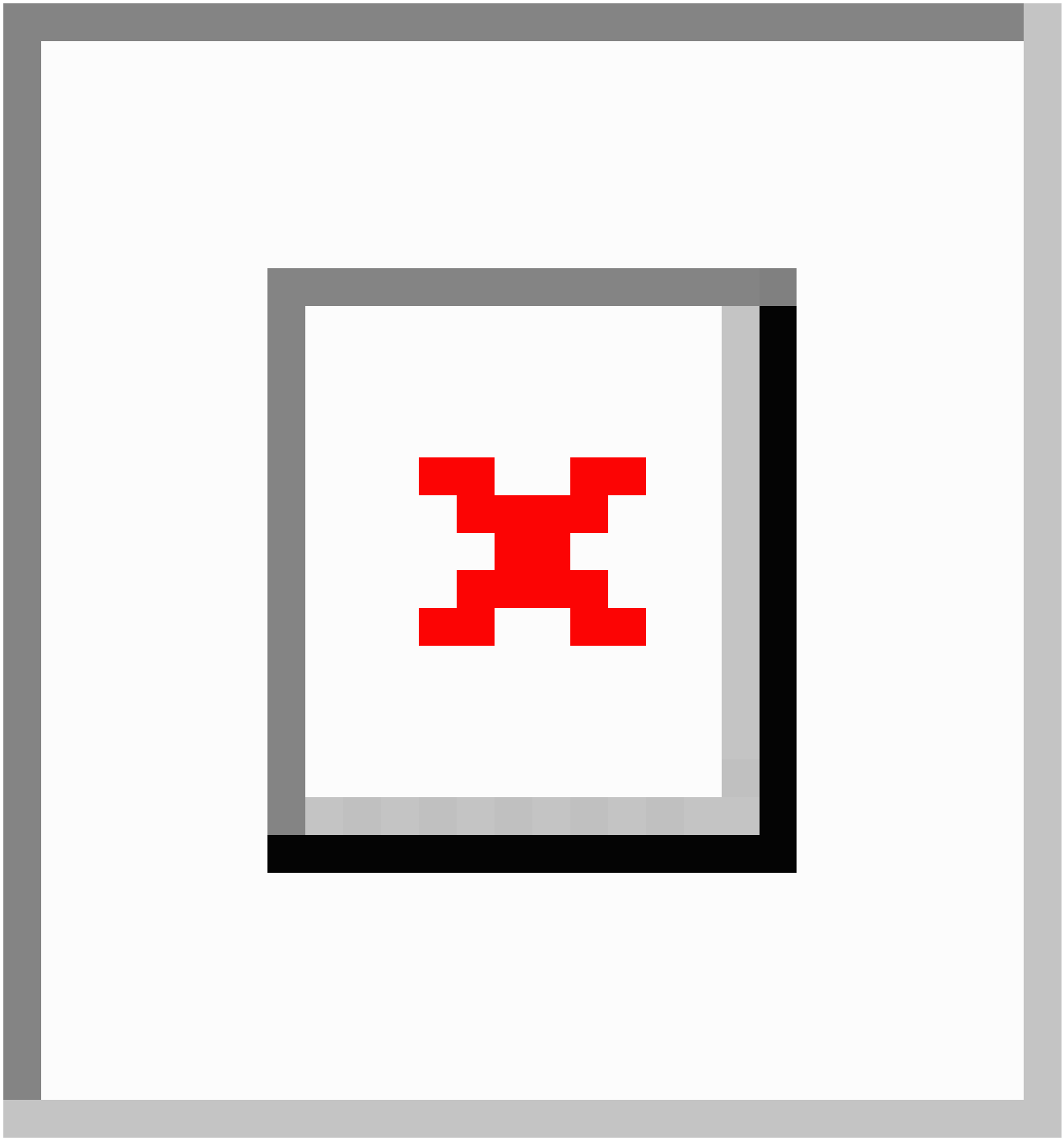
Authorization and Authentication

We designed a simple, yet flexible 3-stage authorization model, which combines permissions for basic object access with permissions regarding the domain of the object to be accessed (with create, read, write, or delete permissions) by a machine or human user of the infrastructure. An overview is provided in Figure 1.

A domain defines the scope of the data managed by the TTP (eg, a research process, a study, a project, or an institute). Multiple domains can be created within a project (eg, to store pseudonyms used in specific subprojects or contexts). Additionally, in gPAS, a domain can have parent and child domains. This results in a tree structure that can be used to tailor permissions to different scopes within individual projects [25].

On the implementation side, we mapped this model to OpenID Connect (OIDC), which is based on OAuth 2.0 [26]. The JavaScript Object Notation Web Token generated in this process contains role names as attributes, which are platform independent and can also be processed on mobile devices. This is important for the additional UIs that we had to develop. As an identity and access management solution, we chose Keycloak, which is in widespread use, has a native administration interface, and is published as open-source software under the Apache License 2.0. Importantly, it can also be connected to a range of directory services usually maintained by hospitals for account and permission management.

Figure 1. Stages of the functional authorization model.



Programmatic Interface

We decided to implement a REST API to extend the programmatic interfaces of E-PIX, gPAS, and gICS and support cross-tool workflows. Due to its stateless nature, this design enables the management and sharing of data across different systems, combined workflows, and calls by external components. One important application of the unified REST API is to combine participant registration with automatic consent checking in gICS, indexing the participant in E-PIX, and generating pseudonyms in gPAS. Furthermore, the REST API can easily be integrated with the developed authentication and authorization model as well as logging and audit trail functionalities. Existing interfaces of MOSAIC tools can also be integrated with the permission model by wrapping them behind REST interfaces.

Graphical Interfaces

Web Interface

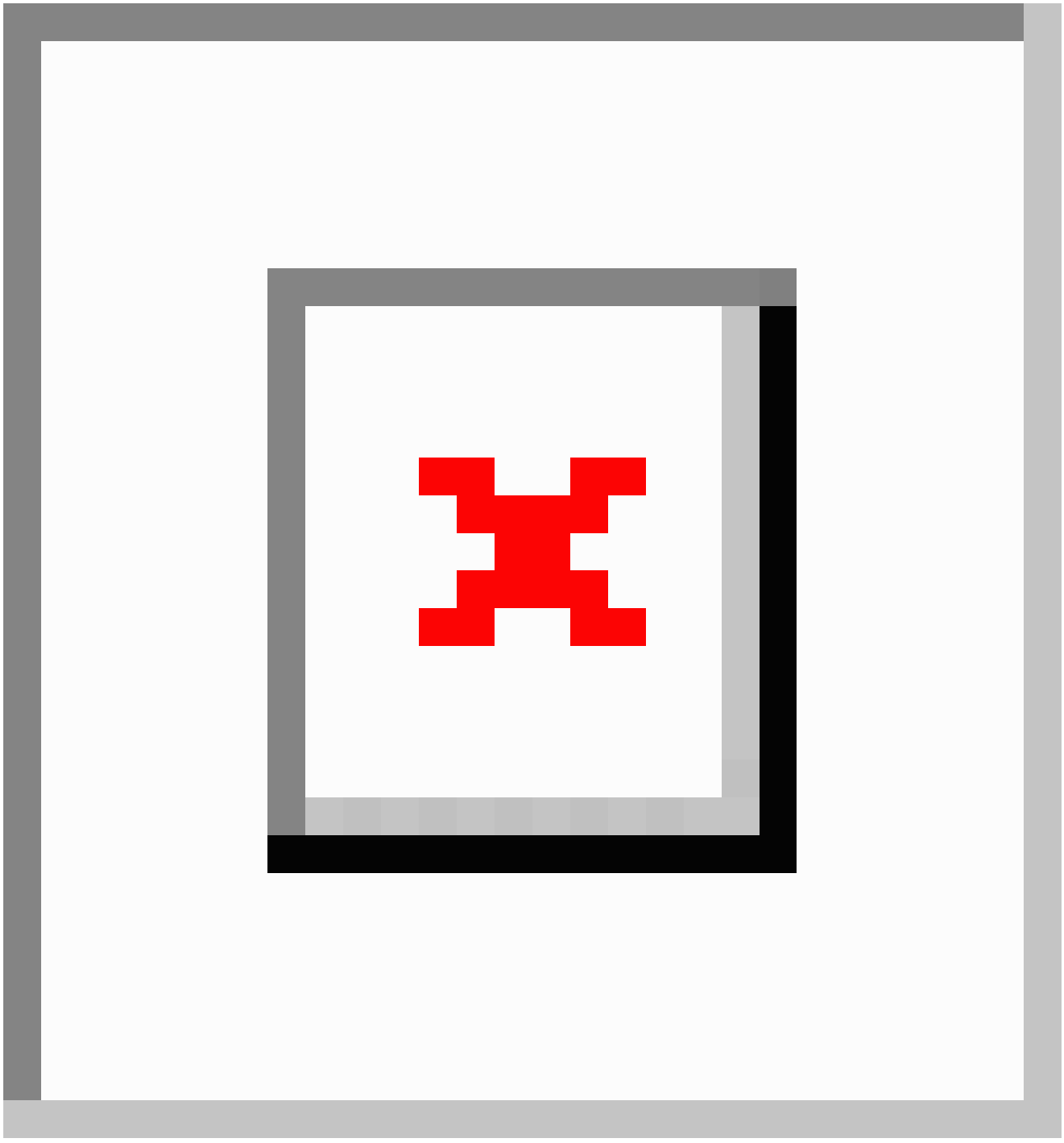
Based on the integrated programmatic API that supports all services, we have also implemented an integrated GUI, which allows accessing all TTP services in a unified manner. Analogously to the programmatic API, the UIs are integrated with the described authentication and authorization model. Users can log into the platform with their account from the connected directory service, which is abstracted way using OIDC with Keycloak. The token generated at log-in contains all assigned permissions, which are used in the UI and sent as a bearer token

with each request to the REST services. A strict content-security-policy workflow blocks the execution of foreign scripts outside the origin domain, thus increasing the level of security. Actions such as participant administration, depseudonymization, or consent administration can be performed through wizards. Users can request essential documents, such as copies of consent, directly from the web application.

Mobile App

The final building block is provided by a mobile app that serves as a direct channel from the TTP services to the participants. The most important application is collecting consent and handling withdrawals. A typical deployment consists of installing the appl on a tablet, which is then configured by study personnel and handed over to the participants (Figure 2).

Figure 2. Workflow of actions in the app.



The study personnel can log into the app using the same log-in data as for the TTP web interface. After the project staff member enters a participant identification code and selects either a consent or a withdrawal form, the selected participant fills out the form. To prevent participants from accessing unauthorized information, the app will be started in kiosk mode. The identification code is either a temporary pseudonym or an already existing pseudonym for the participant, providing direct linkage to the research project managed by the TTP. In the latter case, the app automatically opens the associated consent template. After filling out the form, the participants can enter their name and place of residence, and then, they can put their

signature in a designated field. Afterwards, the staff member provides their signature, confirming that the form has been completed with them as the assigned project staff member.

Supported Pseudonym Algorithms

In our system, generated random numbers are used as pseudonyms. The length is configurable, with a minimum of 6 digits, and is chosen based on the number of pseudonyms that are needed for the respective project. Additionally, we use the Damm algorithm to detect single-digit errors and all adjacent transposition errors with a simple checksum [27]. Moreover, pseudonyms are combined with study- and context-specific prefixes. For example, the pseudonym "BLV-US-123456" could represent an ultrasound ("US") measurement for a study participant in a study called BeLOVE ("BLV"). Finally, our system can also import and manage existing pseudonyms. As those are usually generated using different algorithms and often do not contain a checksum, we mark them as "external" within the system.

Ethical Considerations

This paper covers the design and implementation of a generic research service, which requires no ethics committee approval according to local policies. However, the individual studies that use the service have to apply for ethics approval. For example, the BeLOVE study, which is described as a case study in this paper, was approved by Charité's ethics committee (vote number EA1/066/17).

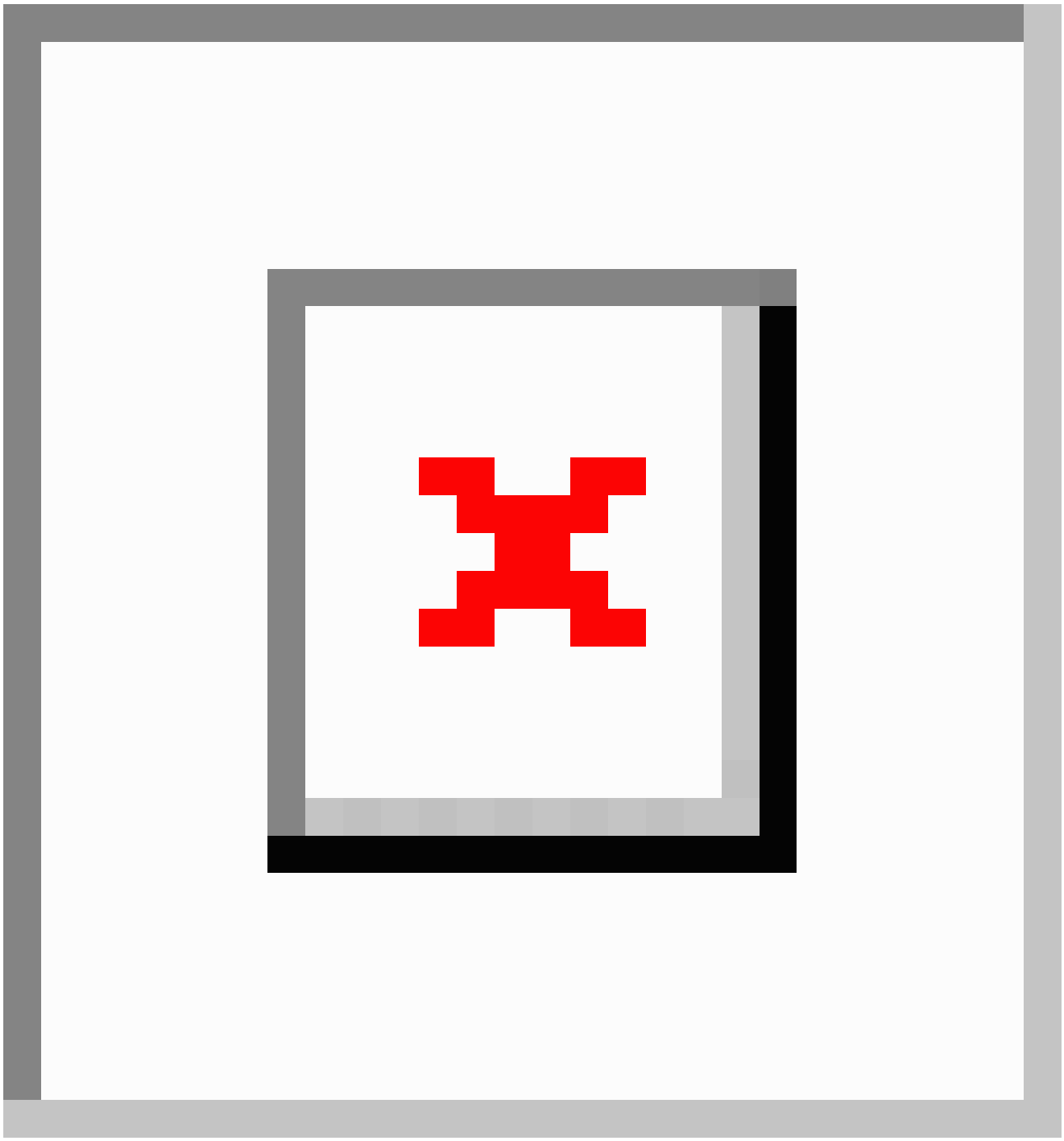
Results

In this section, we will first describe the general architecture of our solution, then cover important implementation details, and finally report on real-world experiences with the platform.

Architecture

The overall architecture is divided into the API, which wraps around the MOSAIC tools, the graphical interfaces oriented toward users, as well as the access and identity management component (Figure 3 presents more details).

Figure 3. Architecture overview, including wrapped MOSAIC stack (core components); systems maintained by the trusted third party (TTP; graphical components as well as access and identity components); systems queried by the TTP (electronic health record [EHR] system and directory services); and systems from which the TTP is queried (Research Electronic Data Capture [REDCap]). E-PIX: Enterprise Identifier Cross-Referencing; gICS: Generic Informed Consent Service; gPAS: Generic Pseudonym Administration Service.



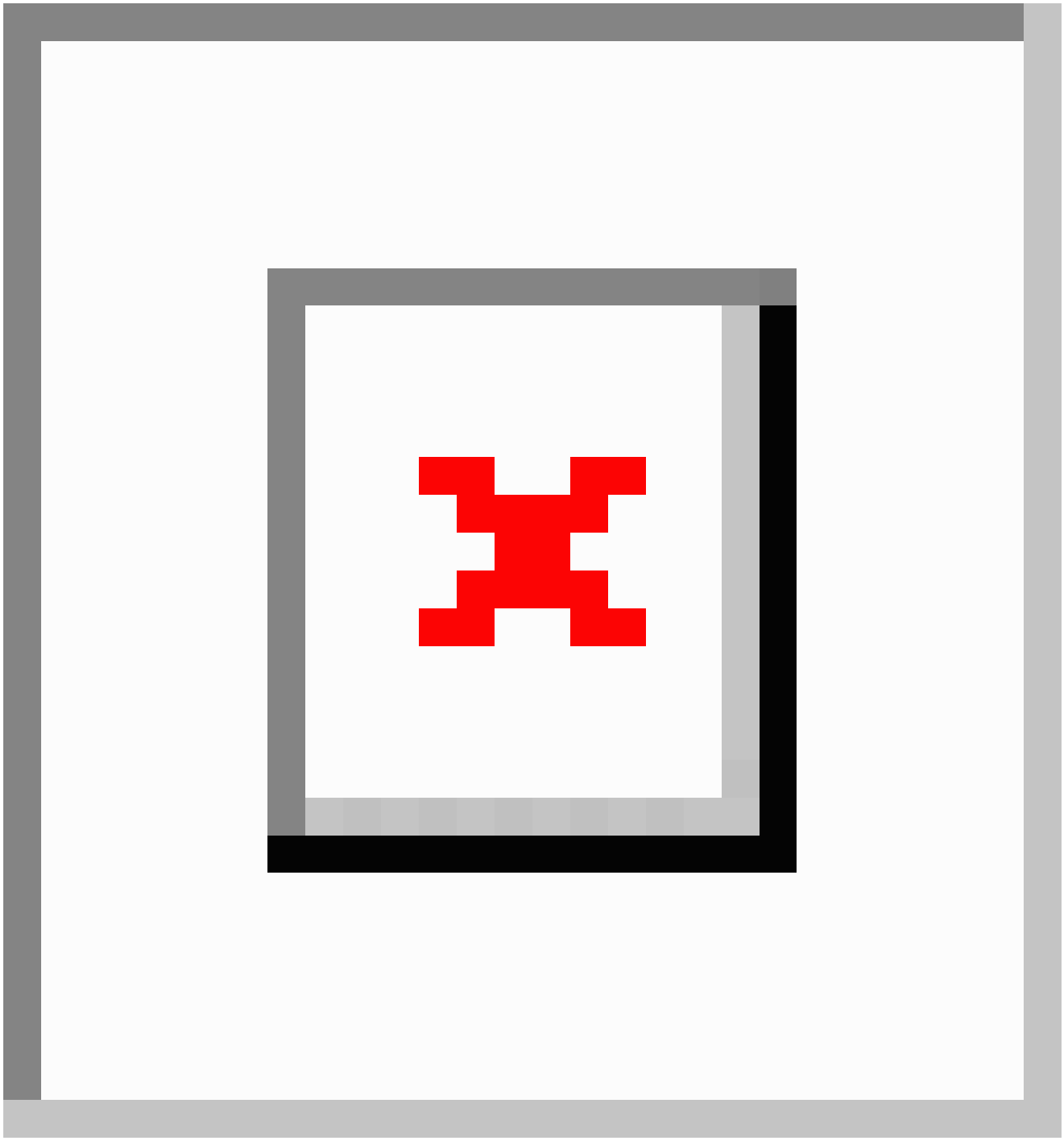
As illustrated, the core components are provided with an interface to the EHR system to support the pseudonymization of patient identities for direct reuse in the respective research context. Other systems that can access the TTP services via the REST API are, for example, electronic data capture systems, such as Research Electronic Data Capture (REDCap), or biobank information systems. All components of the respective interfaces are containerized with Docker [28] and deployed on a Docker swarm [29]. By using OIDC based on OAuth 2.0 as the standard, we were able to integrate other systems via existing packages (eg, Spring-Boot-Security) and allow other applications to access the systems. When modeling the interfaces, we ensured that anything that could be done graphically could also be done programmatically. This keeps the platform open and supports other information systems with the integration of TTP services.

Implementation

The REST API was implemented using Java 13 with the Spring Boot framework [30] by focusing on stable packages, including Spring Security for OIDC, and relying on an established framework. The resulting platform is robust, maintainable, extensible, and flexible. We have implemented 35 generic interfaces so far, most of which are Create-Read-Update-Delete (CRUD) interfaces for the key information objects Domain,

Participant, Identifier, Pseudonym, Consent, and Consent Template (Figure 4), as well as additional directory and search functions for pseudonyms and consents.

Figure 4. Key information objects and their relationships.



The web-based interface (Figures 5 and 6) is implemented using the PHP-based lightweight enterprise web framework Laravel [31]. Laravel uses a Model-View-Controller pattern [32], has a template engine named Blade, and supports agile development processes. By integrating the open-source framework Bootstrap, we were able to implement a responsive front end that could be displayed in browsers on multiple types of devices. The web application directly interfaces with the REST API and does not manage any participant data in a separate database.

Figure 5. Screenshots of the user interface: editing consent information.

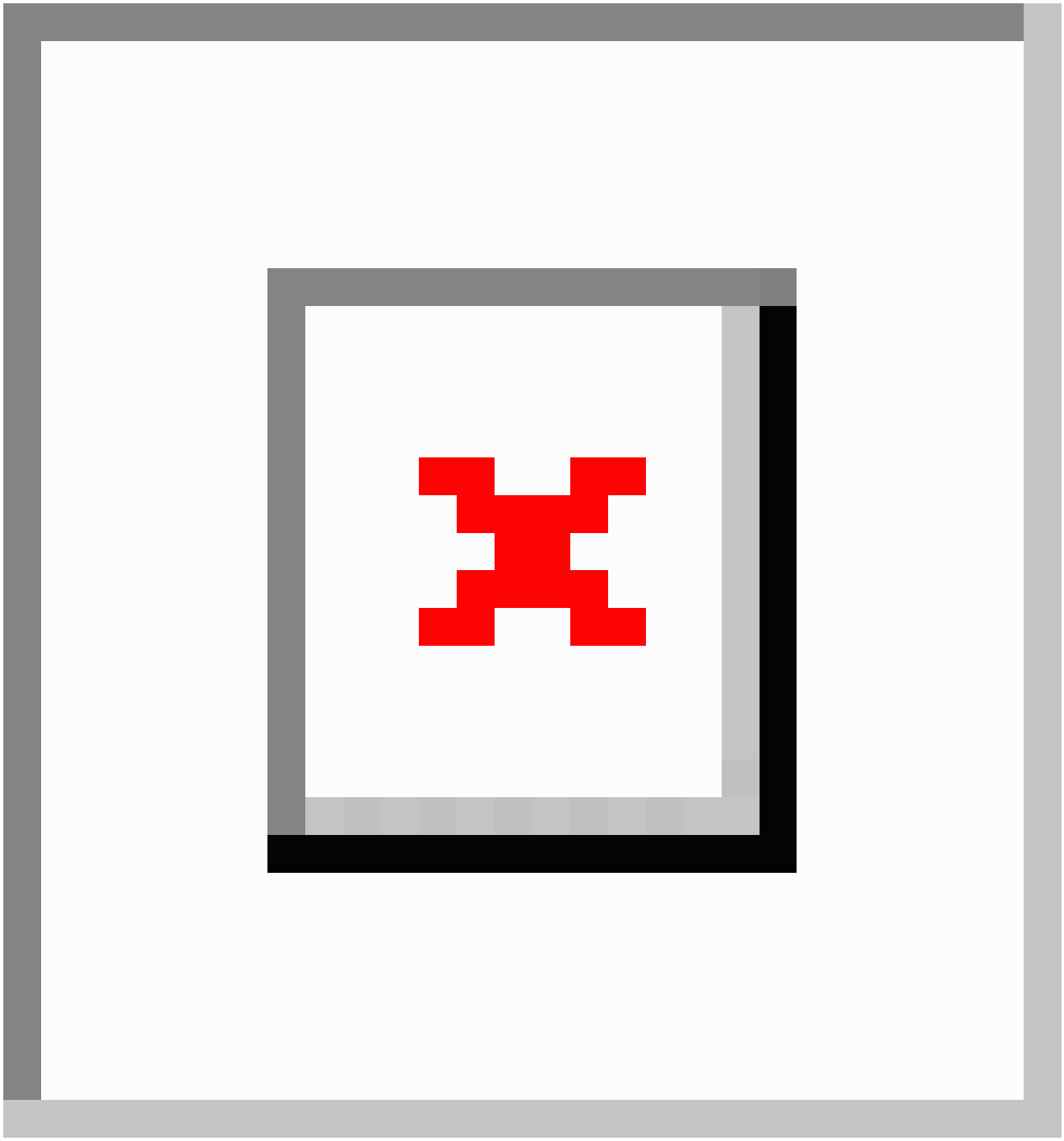
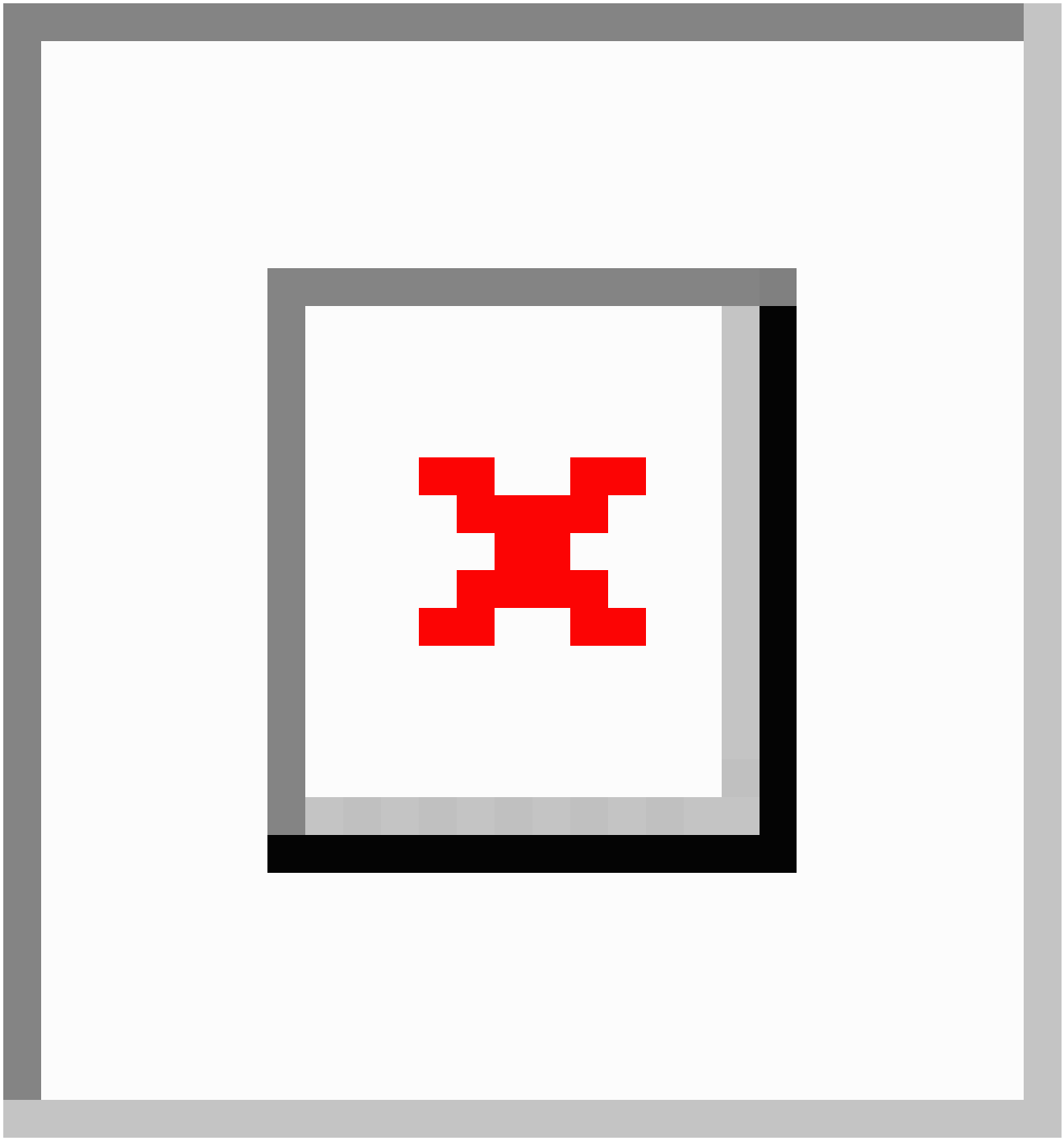


Figure 6. Screenshot of the user interface: overview of consent status.



The app front end (see Figures 7-9) was developed in React Native [33] and then significantly extended to work on tablets integrated into our mobile device management. The application does not permanently store any data on the device, and processing is carried out exclusively via React Native state management.

Figure 7. Screenshot of the consent app: entering or scanning an ID.

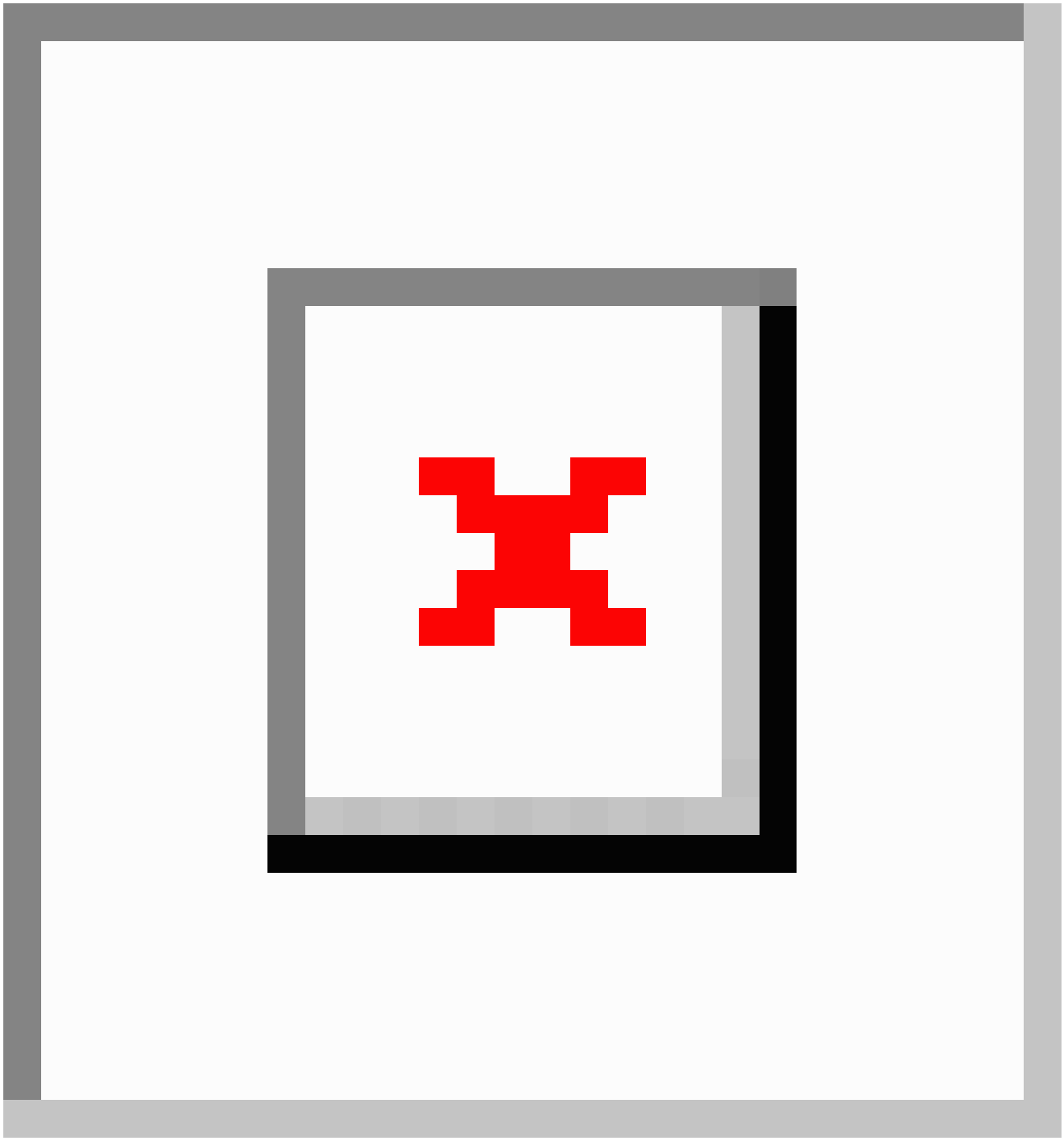


Figure 8. Screenshot of the consent app: filling out consent forms.

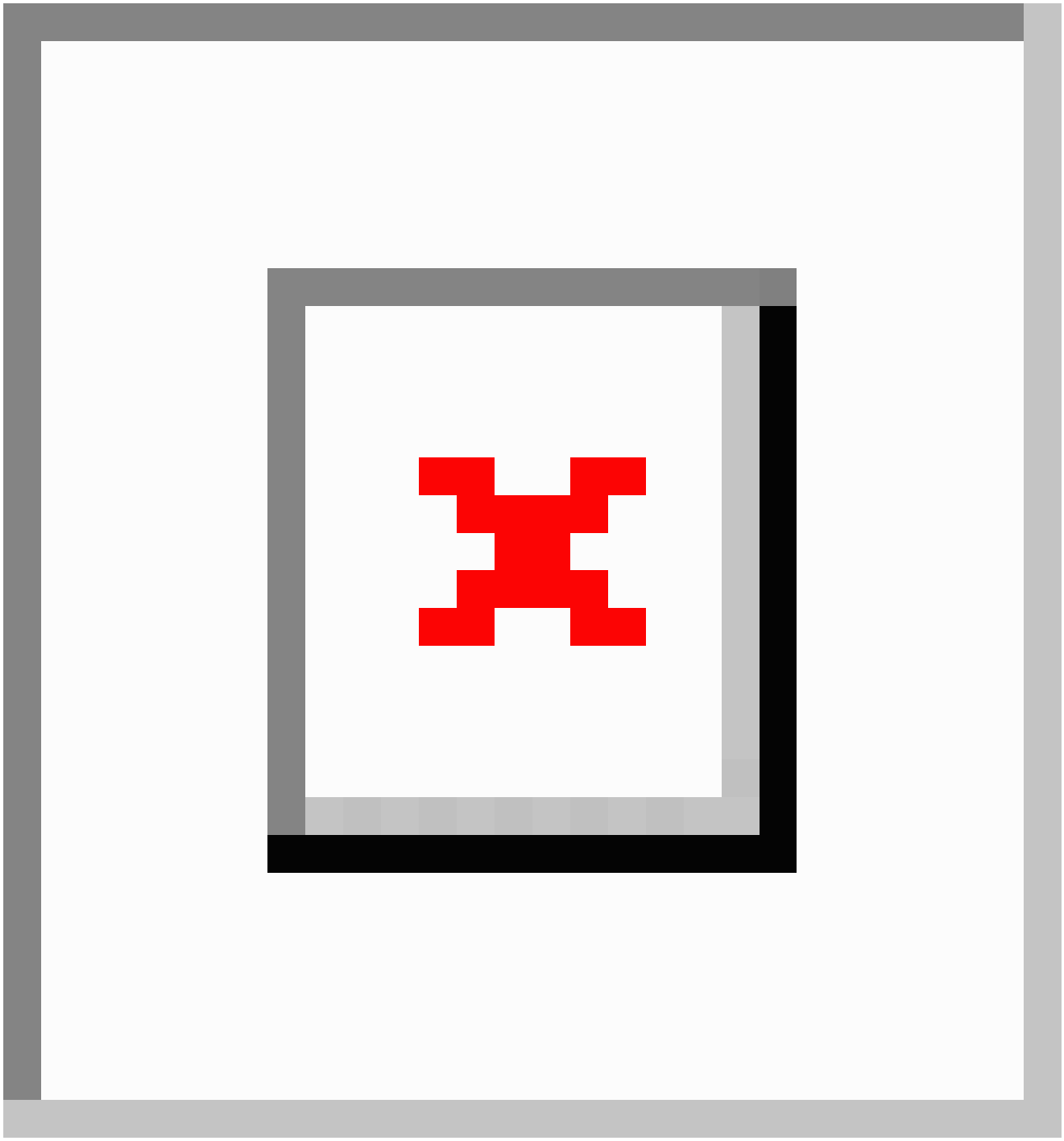
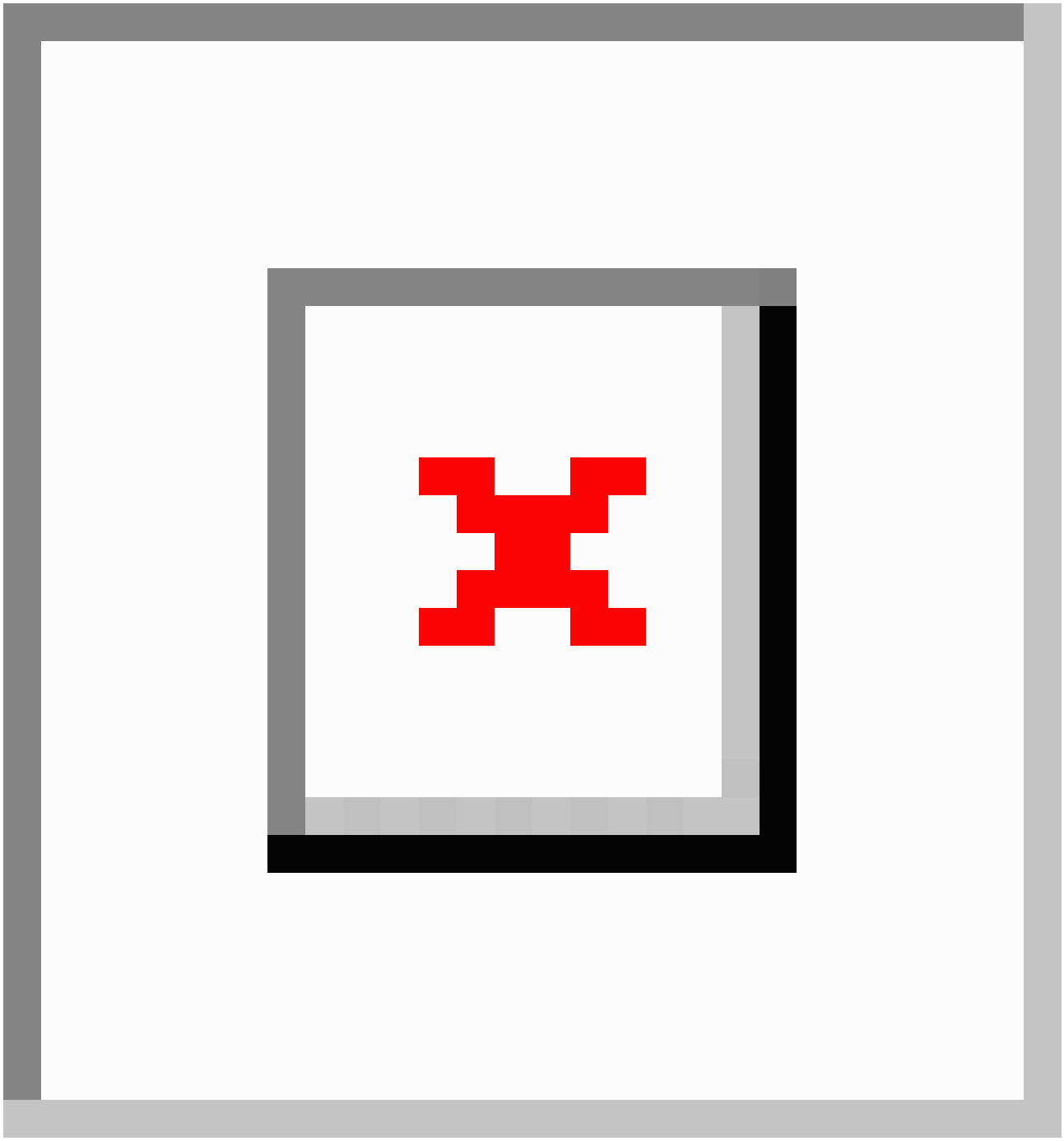


Figure 9. Screenshot of the consent app: sign and submit.



Core Functionalities for Research Projects

As a result of our development efforts, the TTP software stack provides a wide range of functionalities that research projects need. Table 3 provides an overview of frequently used common features.

Table . Essential functionalities provided to research projects.

Component	Process	Description
API[a]	Obtaining a temporary pseudonym	Automated creation of participant placeholders that can be used in third-party systems and later linked to the study identity
App	Electronic consent management	Viewing, completing, saving, and printing an electronic consent template of the respective project under a pseudonym
Web UI[b]	Participant registration	Master data and contact details can be entered manually or imported from the EHR[c] system
Web UI	Participant overview	Provides an overview of the participants and pseudonyms associated with a specific project
API	Integration with other systems	Interface for pseudonymization, depseudonymization, and linkage for third-party systems
Web UI	Depseudonymization	Resolve pseudonym to participant master data
Web UI	Retrieval of usage permissions based on consent information	Retrieve electronic representation of usage permissions from consents associated with a specific patient or participant pseudonym
Web UI	Update participant information	Use pseudonyms to update participant information

[a]API: application programming interface.

[b]UI: user interface.

[c]EHR: electronic health record.

On the API level, these features include integration with other systems to manage pseudonymization, depseudonymization, and data linkage. The app specializes in electronic consent management, specifically viewing, completing, and saving of consent templates. The web-based UI permits registration of participant details; provides an overview of participants, consents, and pseudonyms; supports depseudonymization as well as the retrieval of use permissions based on consent information. CRUD operations for major participant properties and printing consents are also supported.

Experiences in Real-World Operational Settings

The TTP has already supported more than 10 research projects since it was launched in December 2019. As of December 2022, our TTP system manages data of 3610 registered participants with 384,813 pseudonyms and 1762 consent documents. The pseudonyms fall into 2 categories: 40,867 pseudonyms have been assigned to individual participants managed by the TTP and 343,946 pseudonyms to other identifiers (eg, health insurance numbers that are managed by the TTP as part of its support for data linkage). On average, the TTP manages about 11 pseudonyms for each individual participant. As many as 153 research personnel actively engage with the software on a daily basis. Backups of our databases are created every night. These backups are stored for 90 days along with all log files.

As a case study, we will describe how the TTP services are being used by the large-scale BeLOVE study [20], which is carried out as a cooperation between several sites and departments at Charité. BeLOVE uses all services provided, from patient as well as participant registration and consent management, to pseudonym generation for the various diagnostics and phenotyping activities performed during hospitalizations or study visits (about 12 pseudonyms per participant). Compared to the initial planning of the study, which required 2 study staff for the administrative tasks, these staff requirements were in the meantime reduced to zero due to the functionality of our TTP and the associated secure outsourcing of tasks to all study staff. The use of central TTP services has also significantly reduced the efforts required for coordinating BeLOVE and its substudies with the data protection and information security officers. Within Charité's internal data integration platform, consistent pseudonyms and API access to mapping rules are frequently used to link data collected about BeLOVE participants with routine health care data collected during inpatient and outpatient encounters for various types of analyses. Secondary pseudonyms have already been generated for 10 projects in which the data have been analyzed or shared with others.

Discussion

Principal Results

In this paper, we have presented a software stack to support a TTP with its core tasks at a large German academic medical center. Our architecture extends existing systems for key functionalities, identity management, pseudonymization, and consent management with a fine-grained authentication and authorization model, a modern REST API, two types of UIs, and connections to third-party systems. These extensions were necessary to support cross-service workflows on the programmatic as well as the user level and to meet further functional and nonfunctional requirements. Our application is built using various open-source enterprise frameworks and standards (eg, OIDC) to ensure sustainability and integration

with important institutional services (eg, our user directory and leading master patient index). Our experiences with supporting a wide range of research projects with TTP services over a longer period have shown that our approach works and provides functionalities that are generic enough to support a wide range of applications.

Comparison With Prior Work

Our architecture and implementation are based on the MOASIC tools [16], which we have extended with additional components to overcome functional and nonfunctional shortcomings. Most importantly, the publicly available basic versions of the MOSAIC tools are not suitable for handling more complex and flexible workflows with fine-grained authorization. For example, supporting cross-service workflows, like registering a patient, generating pseudonyms, and preparing a consent form as an integrated operation, cannot be implemented without an additional dispatcher component that is currently not publicly available. We solved this by implementing a cross-service REST API. Although the MOSAIC tools already come with an API, it is provided individually for each service and is based on the Simple Object Access Protocol [34], which originates from the IHE web service standards [35] and is complex and slow, requiring managing server-side state. Analogously to an API, the MOSAIC tools also offer GUIs. However, they are provided individually for each service and hence do not enable users to seamlessly perform operations that require interactions with multiple core services. For this reason, we developed a cross-service UI that is based on our API. Additionally, we added functionalities for generating QR codes, versioning consent documents, and starting the system in kiosk mode. Finally, our extensions also improve the system's scalability when executing cross-service operations, such as querying for links between pseudonyms and identifiers, which can be slow when using the MOSAIC tools [36]. We also added comprehensive documentation of administration functions, which is not fully available for the current open-source versions without registration with the vendor [37].

Prior work on TTP-related services usually focused on individual components or algorithms that could support TTP operations, deployments in specific research projects, or high-level architecture overviews.

One well-known example is the one-way hash approach employed by Vanderbilt University Medical Center as part of the ingest process into their deidentified layer within a research data warehouse [38]. Pommering et al [39] describe strategies for how pseudonymization could be used in different contexts, for example, in the secondary use of EHR data or in medical research networks and biobanks. They introduced two models that support repeated depseudonymization as well as one-time use [40]. The former model was later integrated into a concept for sharing large data sets in medical research networks and biobanks [39].

Building on this, Lo Iacono [41] investigated a cryptographic approach for generating consistent pseudonyms in multicentric studies but without describing a specific implementation within a concrete project. Dangl et al [42] describe concepts and requirements for TTP services for a specific biobank of a clinical research group. Heinze et al [43] developed two services based on IHE profiles that have been implemented into the Heidelberg Personal Electronic Health Record. One service is used to capture patient consent, while the other provides a GUI to manage consents. Further components (eg, for pseudonym or identity management) were not described in detail.

Lablans et al [13] introduce the Mainzeliste, which supports managing patient identities and pseudonyms through a web-based front end. Bialke et al [10] introduce the MOSAIC tools, which we also use in our work, as a set of tools supporting central data management for studies or research networks. They also introduce the "dispatcher" as an additional component for building complex workflows [22], which is, as we described above, unfortunately not publicly available.

Aamot et al [44] compare different strategies for depseudonymization in which, among others, the strategy of Pommering et al [39] is compared with alternative approaches. Based on this comparison, they develop a pseudonymization approach using deterministic one-way mappings based on cryptographic protocols. Lautenschläger et al [45] implement and describe a generic and tightly coupled architecture and component for pseudonymization that has been used in several research projects. On the application side, Bahls et al [14] describe a TTP architecture using the MOSAIC tools for the Routine Anonymized Data for Advanced Health Services Research project. Hampf et al [17] benchmark parts of the MOSAIC tools and conclude that it would take several days to register 2 million patients with the hardware setup utilized.

Limitations and Future Work

As the most recent versions of the MOSAIC tools are not distributed as open-source software in a public repository [37], it was not possible for us to make changes to the core tools used. Instead, workarounds had to be implemented at the API or UI level, which is not ideal from an architecture perspective. Moreover, our TTP platform is currently focused on providing intra-institutional services only. In future work, we plan to extend our platform with external interfaces, enabling the TTP to act as a central trustee for multicentric projects. We also aim to implement additional programmatic interfaces following international interoperability standards, in particular, Health Level 7 Fast Healthcare Interoperability Resources [46] and enable study personnel to directly manage the permissions of associated staff. Finally, we plan to introduce a unified pool of consent policy keys to harmonize the permission information that can be queried from our system to enable automated downstream processing that considers consent information.

Conclusions

Scalable and comprehensive TTP services are central to modern data-driven medical research. However, community-based comprehensive platforms that can be used to implement such services are still lacking. We believe that our description of key requirements as well as the insights provided into our flexible architecture that combines core tools with user- and application-oriented workflows and interfaces, including third-party applications, can help other institutions setting up comparable services.

Acknowledgments

The authors would like to thank the BeLOVE (Berlin Longterm Observation of Vascular Events) study team, who have contributed to the improvement of the entire system with their constant feedback. This work was, in part, supported by the German Federal Ministry of Education and Research under grant agreement number 16DTM215 (THS-MED).

Conflicts of Interest

None declared.

References

1. Pommerening K, Sax U, Müller T, Speer R, Ganslandt T, Drepper J. Integrating eHealth and medical research: the TMF data protection scheme. EHealth Comb Health Telemat Telemed Biomed Eng Bioinforma Edge 2008 Jan:5-10 [FREE Full text]
2. Borda A, Gray K, Fu Y. Research data management in health and biomedical citizen science: practices and prospects. JAMIA Open 2019 Dec;3(1):113-125. [doi: 10.1093/jamiaopen/ooz052] [Medline: 32607493]
3. Wang X, Williams C, Liu ZH, Croghan J. Big data management challenges in health research-a literature review. Brief Bioinform 2019 Jan 18;20(1):156-167. [doi: 10.1093/bib/bbx086] [Medline: 28968677]
4. Zhao Z, Chuah JH, Lai KW, et al. Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: a review. Front Comput Neurosci 2023 Feb 6;17:1038636. [doi: 10.3389/fncom.2023.1038636] [Medline: 36814932]
5. Eggert K, Wüllner U, Antony G, et al. Data protection in biomaterial banks for Parkinson's disease research: the model of GEPARD (Gene Bank Parkinson's Disease Germany). Mov Disord 2007 Apr 15;22(5):611-618. [doi: 10.1002/mds.21331] [Medline: 17230444]
6. Bourka A, Drogkaris P, editors. Recommendations on Shaping Technology According to GDPR Provisions - An Overview on Data Pseudonymisation: The European Union Agency for Network and Information Security (ENISA); 2019.
7. Kohlmayer F, Lautenschläger R, Prasser F. Pseudonymization for research data collection: is the juice worth the squeeze? BMC Med Inform Decis Mak 2019 Sep 4;19(1):178. [doi: 10.1186/s12911-019-0905-x] [Medline: 31484555]
8. Pommerening K, Drepper J, Helbing K, Ganslandt T. Leitfaden Zum Datenschutz in Medizinischen Forschungsprojekte: Medizinisch Wissenschaftliche Verlagsgesellschaft (MWV); 2015.
9. Lowrance W. Learning from experience: privacy and the secondary use of data in health research. J Health Serv Res Policy 2003 Jul;8 Suppl 1:S1:2-7. [doi: 10.1258/135581903766468800] [Medline: 12869330]
10. Bialke M, Bahls T, Havemann C, et al. MOSAIC—a modular approach to data management in epidemiological studies. Methods Inf Med 2015;54(4):364-371. [doi: 10.3414/ME14-01-0133] [Medline: 26196494]
11. Geidel L, Bahls T, Hoffmann W. Generische Pseudonymisierung ALS Modul des Zentralen Datenmanagements Medizinischer Forschungsdaten. Universitätsmedizin. 2013. URL: https://www.ths-greifswald.de/wp-content/uploads/2019/09/Poster_DGEpi_PSN_2013_09_27.pdf [accessed 2024-04-10]
12. Rau H, Geidel L, Bialke M, et al. The generic informed consent service gICS: implementation and benefits of a modular consent software tool to master the challenge of electronic consent management in research. J Transl Med 2020 Jul 29;18(1):287. [doi: 10.1186/s12967-020-02457-y] [Medline: 32727514]
13. Lablans M, Borg A, Ückert F. A restful interface to pseudonymization services in modern web applications. BMC Med Inform Decis Mak 2015 Feb 7;15:2. [doi: 10.1186/s12911-014-0123-5] [Medline: 25656224]
14. Bahls T, Pung J, Heinemann S, et al. Designing and piloting a generic research architecture and workflows to unlock German primary care data for secondary use. J Transl Med 2020 Oct 19;18(1):394. [doi: 10.1186/s12967-020-02547-x] [Medline: 33076938]
15. Bruland P, Doods J, Brix T, Dugas M, Storck M. Connecting healthcare and clinical research: workflow optimizations through seamless integration of EHR, pseudonymization services and EDC systems. Int J Med Inform 2018 Nov;119:103-108. [doi: 10.1016/j.ijmedinf.2018.09.007] [Medline: 30342678]
16. Projekte. Unabhängige Treuhandstelle. URL: https://www.ths-greifswald.de/forscher/projekte/ [accessed 2023-08-09]
17. Hampf C, Geidel L, Zerbe N, et al. Assessment of scalability and performance of the record linkage tool E-PIX in managing multi-million patients in research projects at a large university hospital in Germany. J Transl Med 2020 Feb 17;18(1):86. [doi: 10.1186/s12967-020-02257-4] [Medline: 32066455]
18. Unabhängige Treuhandstelle der Universitätsmedizin Greifswald. Universitätsmedizin. URL: https://www.medizin.uni-greifswald.de/de/forschung-lehre/core-units/treuhandstelle/ [accessed 2023-08-09]
19. Siegerink B, Weber J, Ahmadi M, et al. Disease Overarching mechanisms that explain and predict outcome of patients with high cardiovascular risk: rationale and design of the Berlin long-term observation of vascular events (Belove) study. medRxiv 2019 Jul 15:19001024. [doi: 10.1101/19001024]

20. Weber JE, Ahmadi M, Boldt LH, et al. Protocol of the Berlin long-term observation of vascular events (BeLOVE): a prospective cohort study with deep Phenotyping and long-term follow up of cardiovascular high-risk patients. BMJ Open 2023 Oct 31;13(10):e076415. [doi: 10.1136/bmjopen-2023-076415] [Medline: 37907297]
21. Bozoyan C, Fitzer K, Ostrzinski S, et al. Unabhängige Treuhandstelle (THS). NAKO Treuhandstellenkonzept. 2014. URL: https://nako.de/allgemeines/der-verein-nako-e-v/unabhaengig-treuhandstelle/ [accessed 2023-08-09]
22. Bialke M, Penndorf P, Wegner T, et al. A Workflow-driven approach to integrate generic software modules in a trusted third party. J Transl Med 2015 Jun 4;13:176 [FREE Full text]
23. GmbH GG. Das Sollten SIE Über EAN Nummern Wissen. GS1 Germany. URL: https://www.gs1-germany.de/ean-nummern/ [accessed 2024-01-04]
24. 23 patient identifier cross-referencing Hl7 V3 (Pixv3). IHE International. URL: https://profiles.ihe.net/ITI/TF/Volume1/ch-23.html [accessed 2023-09-25]
25. Hampf C, Bialke M. Unabhängige Treuhandstelle der Universitätsmedizin Greifswald. gPAS Anwenderhandbuch. 2023. URL: https://www.ths-greifswald.de/gpas/handbuch
26. Ma W, Sartipi K, Sharghigoorabi H, Koff D, Bak P. Openid connect as a security service in cloud-based medical imaging systems. J Med Imaging (Bellingham) 2016 Apr;3(2):026501. [doi: 10.1117/1.JMI.3.2.026501] [Medline: 27340682]
27. Damm MH. Total Anti-Symmetrische Quasigruppen [article in German].: Philipps-Universität Marburg; 2004 URL: https://archiv.ub.uni-marburg.de/diss/z2004/0516/ [accessed 2024-04-10]
28. Docker overview. Docker Docs. 2023. URL: https://docs.docker.com/get-started/overview/ [accessed 2023-08-09]
29. Docker swarm overview. Docker Docs. 2023. URL: https://docs.docker.com/engine/swarm/ [accessed 2023-10-09]
30. Spring Boot. URL: https://spring.io/projects/spring-boot/ [accessed 2023-08-14]
31. The PHP framework for web artisans. Laravel. URL: https://laravel.com/ [accessed 2023-08-14]
32. Krasner G, Pope S. A cookbook for using the model-view controller user interface paradigm in Smalltalk-80. JOOP 1988 Jan [FREE Full text]
33. Kopp M. Entwicklung Einer App Zur Erfassung von Einverständniserklärungen Zur Datenverarbeitung Im Rahmen Einer Medizinischen Studie an Der Charité Berlin: HTW Berlin; 2021.
34. SOAP version 1.2 part 1: messaging framework (second edition). W3. URL: https://www.w3.org/TR/soap12/ [accessed 2023-08-10]
35. Appendix V: web services for IHE transactions. URL: https://profiles.ihe.net/ITI/TF/Volume2/ch-V.html [accessed 2023-09-25]
36. Fischer H, Röhrig R, Thiemann VS. A generic IT infrastructure for identity management and pseudonymization in small research projects with heterogeneous and distributed data sources under consideration of the GDPR. Stud Health Technol Inf 2019 Aug 21;264:1837-1838. [doi: 10.3233/shti190673]
37. Community. Unabhängige Treuhandstelle. URL: https://www.ths-greifswald.de/forscher/community/#collapse-1-5454 [accessed 2023-08-11]
38. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform 2014 Dec;52:28-35. [doi: 10.1016/j.jbi.2014.02.003] [Medline: 24534443]
39. Pommerening K, Schröder M, Petrov D, Schlösser-Faßbender M, Semler SC, Drepper J. Pseudonymization service and data custodians in medical research networks- and biobanks. : Gesellschaft für Informatik eV; 2006. URL: https://dl.gi.de/handle/20.500.12116/23646 [accessed 2023-08-09]
40. Pommerening K, Reng M. Secondary use of the EHR via pseudonymisation. Stud Health Technol Inform 2004;103:441-446. [Medline: 15747953]
41. Lo Iacono L. Multi-centric universal pseudonymisation for secondary use of the EHR. Stud Health Technol Inform 2007;126:239-247. [Medline: 17476066]
42. Dangl A, Demiroglu SY, Gaedcke J, et al. The IT-infrastructure of a biobank for an academic medical center. Stud Health Technol Inform 2010;160(Pt 2):1334-1338. [Medline: 20841901]
43. Heinze O, Birkle M, Köster L, Bergh B. Architecture of a consent management suite and integration into IHE-based regional health information networks. BMC Med Inform Decis Mak 2011 Oct 4;11:58. [doi: 10.1186/1472-6947-11-58] [Medline: 21970788]
44. Aamot H, Kohl CD, Richter D, Knaup-Gregori P. Pseudonymization of patient Identifiers for translational research. BMC Med Inform Decis Mak 2013 Jul 24;13(1):75. [doi: 10.1186/1472-6947-13-75] [Medline: 23883409]
45. Lautenschläger R, Kohlmayer F, Prasser F, Kuhn KA. A generic solution for web-based management of pseudonymized data. BMC Med Inform Decis Mak 2015 Nov 30;15:100. [doi: 10.1186/s12911-015-0222-y] [Medline: 26621059]
46. HL7 FHIR. URL: https://www.hl7.org/fhir/ [accessed 2023-08-09]

Abbreviations

API: application programming interface
BeLOVE: Berlin Longterm Observation of Vascular Events
BLV: BeLOVE

CRUD: Create-Read-Update-Delete
E-PIX: Enterprise Identifier Cross-Referencing
EHR: Electronic Health Record
gICS: Generic Informed Consent Service
gPAS: Generic Pseudonym Administration Service
GUI: Graphical user interface
IHE: Integrating the Healthcare Enterprise
NAKO: German National Cohort
OIDC: OpenID Connect
PHP: Hypertext Preprocessor
PIX: Patient Identifier Cross-Reference
REDCap: Research Electronic Data Capture
REST: representational state transfer
TMF: Technology, Methods, and Infrastructure for Networked Medical Research
TTP: trusted third party

Edited by A Benis; submitted 25.09.23; peer-reviewed by HJ Kim, X Wu; revised version received 15.02.24; accepted 17.02.24; published 17.04.24.

Please cite as:
Wündisch E, Hufnagl P, Brunecker P, Meier zu Ummeln S, Träger S, Kopp M, Prasser F, Weber J
Development of a Trusted Third Party at a Large University Hospital: Design and Implementation Study
JMIR Med Inform 2024;12:e53075
URL: https://medinform.jmir.org/2024/1/e53075
doi:10.2196/53075

Original Paper

Interpretable Deep Learning System for Identifying Critical Patients Through the Prediction of Triage Level, Hospitalization, and Length of Stay: Prospective Study

Yu-Ting Lin[1*], MSc; Yuan-Xiang Deng[1*], MSc; Chu-Lin Tsai[2*], MD, SCD; Chien-Hua Huang[2*], MD, PhD; Li-Chen Fu[1*], PhD

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

[2]Department of Emergency Medicine, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan

[*]all authors contributed equally

Corresponding Author:
Li-Chen Fu, PhD
Department of Computer Science and Information Engineering
National Taiwan University
CSIE Der Tian Hall No. 1, Sec. 4, Roosevelt Road
Taipei, 106319
Taiwan
Phone: 886 935545846
Email: lichen@ntu.edu.tw

Abstract

Background: Triage is the process of accurately assessing patients' symptoms and providing them with proper clinical treatment in the emergency department (ED). While many countries have developed their triage process to stratify patients' clinical severity and thus distribute medical resources, there are still some limitations of the current triage process. Since the triage level is mainly identified by experienced nurses based on a mix of subjective and objective criteria, mis-triage often occurs in the ED. It can not only cause adverse effects on patients, but also impose an undue burden on the health care delivery system.

Objective: Our study aimed to design a prediction system based on triage information, including demographics, vital signs, and chief complaints. The proposed system can not only handle heterogeneous data, including tabular data and free-text data, but also provide interpretability for better acceptance by the ED staff in the hospital.

Methods: In this study, we proposed a system comprising 3 subsystems, with each of them handling a single task, including triage level prediction, hospitalization prediction, and length of stay prediction. We used a large amount of retrospective data to pretrain the model, and then, we fine-tuned the model on a prospective data set with a golden label. The proposed deep learning framework was built with TabNet and MacBERT (Chinese version of bidirectional encoder representations from transformers [BERT]).

Results: The performance of our proposed model was evaluated on data collected from the National Taiwan University Hospital (901 patients were included). The model achieved promising results on the collected data set, with accuracy values of 63%, 82%, and 71% for triage level prediction, hospitalization prediction, and length of stay prediction, respectively.

Conclusions: Our system improved the prediction of 3 different medical outcomes when compared with other machine learning methods. With the pretrained vital sign encoder and repretrained mask language modeling MacBERT encoder, our multimodality model can provide a deeper insight into the characteristics of electronic health records. Additionally, by providing interpretability, we believe that the proposed system can assist nursing staff and physicians in taking appropriate medical decisions.

(JMIR Med Inform 2024;12:e48862) doi:10.2196/48862

KEYWORDS

emergency department; triage system; hospital admission; length of stay; multimodal integration

Introduction

Background

Emergency services are an essential aspect of the health care system in hospitals, and the demand for these services has increased exponentially in recent years. For instance, due to a rising number of elderly patients, a high volume of low-acuity patients waiting for the emergency department (ED), and limited access to medical resources in the community, it may take a long time for patients to receive medical treatment in the ED. Additionally, the situation has worsened with the shortage of experienced health care providers. In the ED, this can cause many severe clinical outcomes, such as delayed diagnosis, longer patient wait times, and increased mortality rates. Moreover, the patient and the standard health care operation procedure may be disturbed. Therefore, prioritizing ED visits and maintaining the regular operation of the health care system are essential.

Triage is the process of accurately assessing patients' symptoms and providing them with proper clinical treatment in the ED. Patients are assigned different priorities depending on their vital signs and chief complaints, and the judgment description from the nursing staff [1]. Many countries have developed their triage process to stratify the clinical severity of patients and thus distribute medical resources. For instance, the US Emergency Severity Index (ESI), Canadian Triage and Acuity Scale (CTAS) [2], and Taiwan Triage Acuity Scale (TTAS) are designed to improve the triage prioritizing process [3-5]. In terms of personnel, hospitals employ dedicated nurses who have been certified by the authorities to undertake the triage process. It is also essential to maintain the quality of education, training, and evaluation of those professionals, which is more difficult nowadays with the increase in the complexity of emergency care and the increase in the number of patients visiting the ED nationwide [6]. Although many standardized scales have been adopted to improve the process, there are still some limitations of the current triage system [7-9]. Among these issues, the lack of capability to prioritize patients and assign patients to appropriate triage levels is the most serious problem. According to records collected in Taiwan from 2009 to 2015, 167,598 out of 268,716 (nearly 60%) visits in the ED were assigned to level 3 in the triage process. In addition, 5-level triage mainly relies on an experienced nurse's diagnosis that is based on a mix of subjective and objective criteria. Any human judgement errors or even inaccurate measurements that occur during the triage assessment can severely affect the outcome.

Related Work

Contextualized Word Embedding

A word vector is an attempt to mathematically capture the syntactic and semantic features of a word and represent its meaning simultaneously. Computers calculate how often words appear next to each other by going through a large corpus. For instance, with GloVe [10] or word2vector [11], the word can be projected into a high-dimensional vector for further tasks.

Although these traditional word embedding methods are easy to understand and simple to implement, some limitations still need to be addressed. For example, after applying word vectors, it would be tough to train systems equipped with the softmax function owing to a large number of categories. On the other hand, the GloVe word embedding involves a numeric representation of a word regardless of where the word occurs in the sentence and the different meanings the word may have. Hence, several language models have been proposed to address these limitations, including embeddings from language models (ELMo) [12], bidirectional encoder representations from transformers (BERT) [13], and generative pretrained transformer (GPT) [14]. These celebrated language models generate general contextualized sentence embeddings by using a large scale of unlabeled corpora.

Among these famous models, BERT is the most popular model commonly used in solving natural language processing (NLP) tasks. BERT is a language model trained bidirectionally, which means that as compared to single-direction language models, it can provide a more profound sense of language context and flow. Moreover, instead of predicting the next word in the sentence, BERT also uses a novel method called "mask language modeling" (MLM). This novel algorithm randomly masks the words and then predicts them. BERT relies on the transformer architecture; however, since BERT aims to generate a language representation model, it only uses the transformer encoder by stacking them up. Later, with the help of MLM and "next sentence prediction" (NSP), BERT can achieve significant performance on lots of NLP downstream tasks by further fine-tuning on specific domains.

Deep Learning for Tabular Data

In statistics, tabular data refer to data organized in a table. Within the table, the rows and columns represent observations and attributes for those observations, respectively. Although many domains like vision, NLP, and speech enjoy the benefit of deep learning models, tabular data using deep learning methods remain questionable. On the other hand, when it comes to handling tabular data, the traditional machine learning method dominates most of the benchmarks and is commonly used in competitions, such as Kaggle, around the world. The conventional machine learning methods include methods based on decision tree (DT) such as extreme gradient boosting (XGBoost) [15], category boosting (CatBoost) [16], and light gradient boosting machine (LightBGM) [17]. The strength of these DT-based methods is that their output is easy to understand and available to provide interpretability without requiring any statistical knowledge. However, there are still some limitations of DT-based methods. Among these limitations, the most serious is that DT-based methods do not allow efficient learning with image or text encoders. Hence, many experts turn to deep learning methods instead of DT-based methods. Deep learning models enable end-to-end learning for tabular data and have many benefits at the same time. First, they can achieve better performance in a bigger data set. Second, they can alleviate the need for feature engineering. Finally, they encode multiple data types efficiently, like images along with tabular data.

However, the shortcoming of most deep learning methods is that they cannot provide interpretability. Fortunately, researchers have been aware of the problem in recent years, and several

deep learning models with interpretability have been proposed, such as TabNet [18], neural oblivious decision ensembles (NODE) [19], and TabTransformer [20].

Current Work in the Triage System

Although current triage systems, such as the ESI and TTAS, follow clear guidelines to assign patient acuity, it implicitly leaves room for clinician interpretation. Hence, the diagnosis still depends heavily on the judgment and experience of individual nursing staff. Several studies have shown that cognitive biases can influence clinical judgments [6]. In written case scenarios at multiple EDs, the average accuracies of nurses were 56.2%, 59.2%, and 59.6% in Taiwan, Brazil, and Switzerland, respectively [21]. In view of this, some studies [6,21,22] have turned to the use of artificial intelligence (AI) systems to assist with decision-making in triage. They also demonstrate the system's effectiveness with higher accuracy from the assisted means.

Numerous studies have attempted to use traditional machine learning methods in their approaches. Choi et al [6] used 3 types of conventional machine learning methods, including logistic regression, random forest, and XGBoost, to predict the Korea Triage Acuity Scale (KTAS) level. They used patients' chief complaints as categorical features, meaning that they assigned a key code to each symptom. Their best model using random forest achieved precision, recall, and area under the receiver operating characteristic curve values of 0.737, 0.730, and 0.917, respectively. Liu et al [22] used CatBoost as their model; however, the study focused on distinguishing the mis-triage of patients in levels 3 and 4 since they believed that the under-triage of critically ill patients could be life-threatening. Their model was able to reduce the life-threatening mis-triage rate from 1.2% to 0.9% prospectively. Ivanov et al [21] carried out a series of experiments to demonstrate the effectiveness of their novel idea "clinical natural language processing (C-NLP)." To cope with free-text data, C-NLP uses sentence tokenization, word tokenization, and part-of-speech tagging to extract the meaning behind free-text data. Their best model included C-NLP and XGBoost, and it was able to achieve an accuracy of 75.7%, which is 26.9% higher than the average nurse's accuracy.

The previously mentioned studies [6,21,22] achieved great performance in dealing with triage-level problems; however, these methods still have some limitations. Our proposed model aims to address these limitations and alleviate them. Multimedia Appendix 1 presents comparisons between earlier work and our study in different aspects.

Goal of This Study

Although the studies mentioned in the previous section successfully demonstrated that AI improved the triage system for predicting triage level, they unfortunately had some serious drawbacks. In this study, we attempted to overcome these drawbacks while developing an appropriate prediction system based on triage information, including demographics, vital signs, and chief complaints. We propose a system that can handle the collected heterogeneous data, including tabular data and free-text data. The proposed system is capable of providing precise suggestions for ED staff in hospitals, and it has interpretability for better acceptance by users. Moreover, it is applicable to real-world situations.

Methods

System Overview

In this study, we have proposed a system comprising 3 subsystems, with each of them handling 1 task. As shown in Figure 1, these tasks include triage level prediction, hospitalization prediction, and length of stay prediction, which are important outcomes in the ED of a hospital. Since these subsystems are developed in a similar training process, we will first introduce the conceptual level of the typical training process of each model in each subsystem and then provide further information. Finally, we will show the detailed design of each model in each subsystem.

Our study focuses on establishing an effective and precise AI system to predict the criticality of patients waiting in the ED of hospitals. By leveraging a model trained on a data set where data labels include different scales, we look forward to developing a robust model that can provide more information to the physician and nursing staff. Moreover, to assist them in making precise medical decisions, our proposed system offers multiple prediction outcomes, including triage-level classification, hospitalization estimation, and length of stay.

The system flowchart is shown in Figure 2. The system can be divided into 3 stages: pretraining stage, fine-tuning stage, and testing stage. Additionally, 2 data sets were used in our study. One was the National Taiwan University Hospital (NTUH) retrospective data set, and the other was the NTUH prospective data set collected from May 26, 2020, till February 21, 2022. These 2 data sets will be elaborated in the following sections.

In the pretraining stage, a large amount of retrospective data were used to pretrain the encoders to learn the basic information of the medical data. In addition, the pretrained encoders were transferred to the second stage. In the fine-tuning stage, we used prospective data with golden labels to fine-tune the pretrained encoder. Therefore, when the diagnosis outcomes from the physician are treated as the ground truth label, the model is more applicable to real-world situations. Finally, in the testing stage, we implemented our system in the hospital and assessed the effectiveness of the system.

Figure 1. The proposed system comprising 3 subsystems that are responsible for different tasks. AI: artificial intelligence.

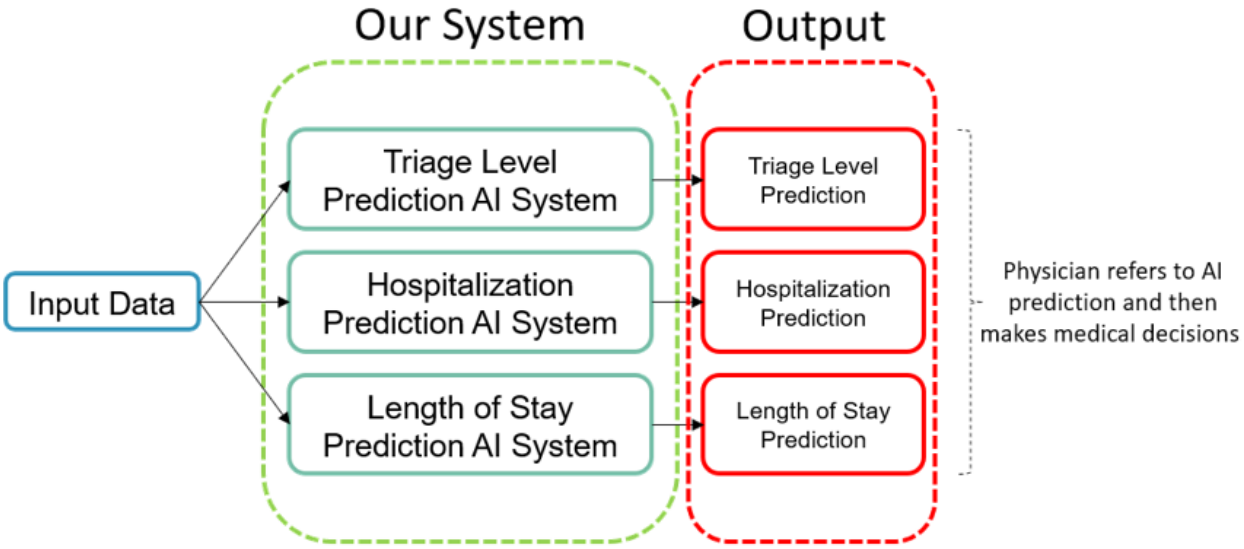
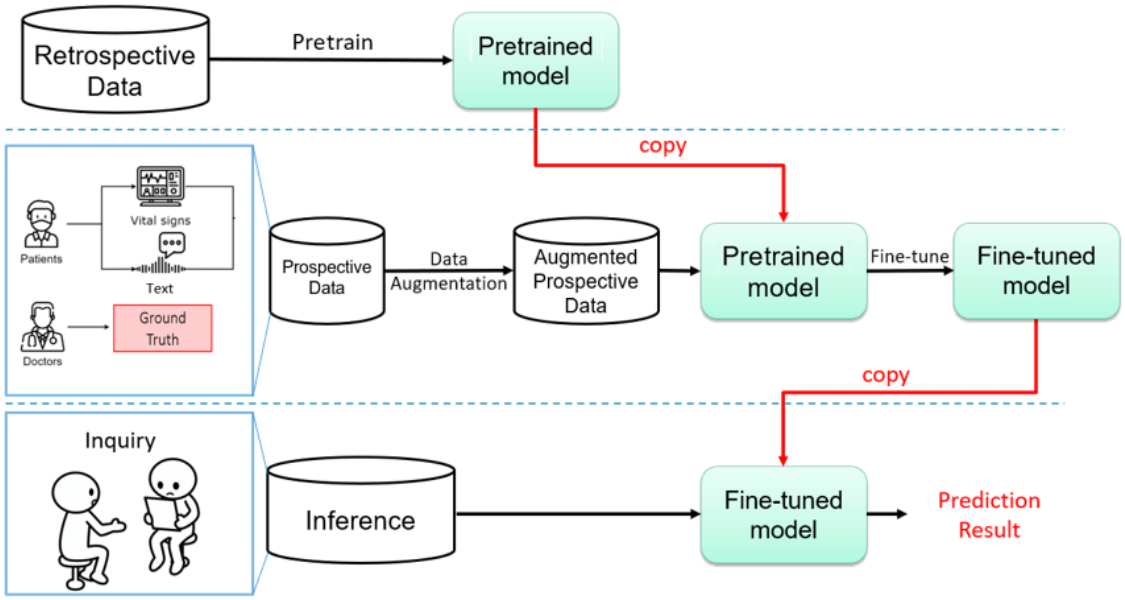


Figure 2. System flowchart.



Ethical Considerations

This study has been approved by the NTUH Institutional Review Board (201606072RINA, 201911054RINA, 202108090RINC).

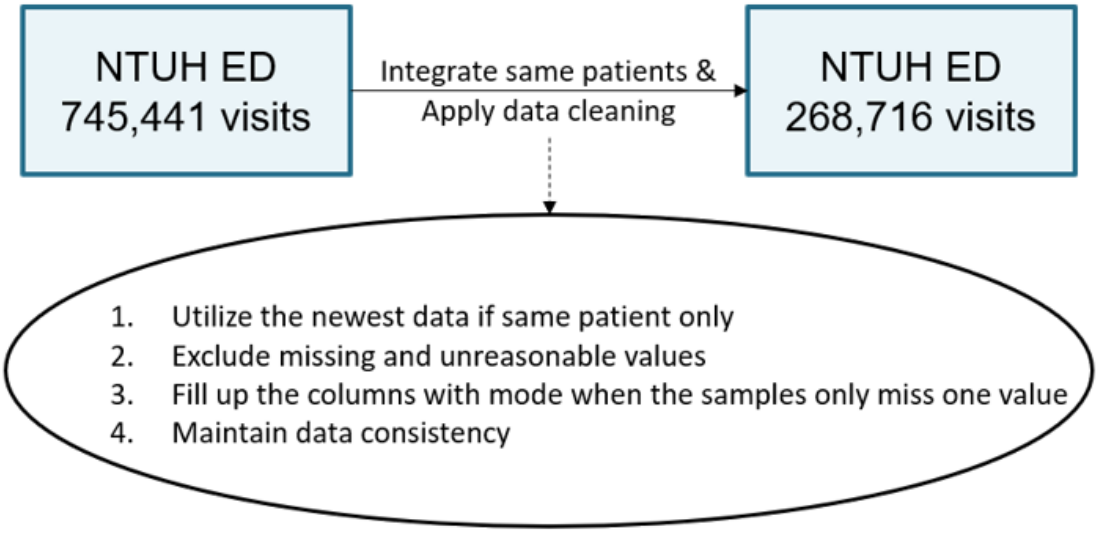
Data Preparation

NTUH Retrospective Data Set

The NTUH is a tertiary academic medical center that has almost 2400 beds and 100,000 emergency room visits per year. After receiving approval from the NTUH Institutional Review Board, we obtained the NTUH retrospective data set, which contained a total of 745,441 electronic health records (EHRs) of patients who visited the ED from the years 2009 to 2015. Since triage is the starting point of care for the ED, it is essential to ensure consistent and precise estimation of patients. The records were evaluated by dedicated personnel who were certified by the Taiwan Union of Nurses Association (TUNA), following a standard protocol.

As shown in Figure 2, in the first stage, we used the NTUH retrospective data set to pretrain our model. However, in the NTUH retrospective data set, we needed to unify the uncleaned data (Multimedia Appendix 2) initially as the members of the nursing staff have their own ways to record the estimation. We included all patients aged 20 years or older who attended the ED and excluded patients whose EHR data contained missing or unreasonable values. Unreasonable data had unreasonable values, which may have resulted from typing errors. For instance, the diastolic pressure and systolic pressure may be typed in reverse, or a nurse may accidentally omit a digit when entering values on the computer. In such a scenario, even though we may be able to infer the original intended values by examining individual data, we cannot consider this a correct sample for use. After data cleaning and merging, only 268,716 patients were enrolled in our program (Figure 3).

Figure 3. Preprocessing of the National Taiwan University Hospital (NTUH) retrospective data set. ED: emergency department.



NTUH Prospective Data Set

Each patient who visits the ED will have a PDF document form generated (triage examination and evaluation record). These records are kept for the physician to make a diagnosis. The records comprise 2 types of information. The first is structural data, including patient demographics, triage information, and vital signs, and the second is textual data, including chief complaints, historical medical information, and drug allergy.

In general, it is impossible to directly use the aforementioned records to train the model, and thus, data preprocessing is needed to extract the data from the records. We used the PDFMiner library in Python code to extract the information from the document forms as "structural data" and applied a transformation function to generate "textual data."

The information extracted from the forms and records can be divided into 2 groups: target prediction and patient feature. Detailed explanations of the patient features are provided in Table 1. On the other hand, the target ground truth contains 3 different tasks. The first task is triage level prediction, which is a 4-class classification problem, where the physician's suggestion is considered (golden standard label that is obtained from the physician by observing the process of patient diagnosis) instead of the traditional triage level. A lower level indicates that the patient more urgently requires immediate attention. The second task is hospitalization prediction, which is a 2-class classification problem, where "0" represents that the patient needs to be discharged by the hospital and "1" represents that the patient needs to be admitted. The last task is length of stay, which is a 3-class classification problem, where "0" represents that the patient will stay in the ED for less than 6 hours, "1" represents that the patient will stay in the ED for 6 to 24 hours, and "2" represents that the patient will stay in the ED for more than 24 hours.

Table 1. Detailed explanation of structural variables.

Variable	Explanation
Demographics	
Age	Patient age
Sex	Patient gender
Triage information	
Session	Patient arrival time
Return in 24 hours	Number of times the patient revisited the ED[a] in 24 hours
Clinic visit mode	Patient arrival mode
Work related	Whether the patient visited the ED because of a work accident
On the way to work	Whether the patient was on the way to work before visiting the ED
Vital sign information	
Systolic pressure	Systolic blood pressure
Diastolic pressure	Diastolic blood pressure
Pulse	Pulse
Oxygen	Oxygen saturation
Respiration	Respiration
Body temperature	Body temperature
Acute change	Any acute changes before entering the ED
Fever	Whether the patient has fever
Pain index	Self-evaluated pain score
GCS-E	Glasgow Coma Scale score of the patient (eye opening)
GCS-V	Glasgow Coma Scale score of the patient (verbal response)
GCS-M	Glasgow Coma Scale score of the patient (motor response)
Major disease	Whether the patient has an IC[b] card for severe illness
Admission count	The number of times the patient went to the hospital in 1 year
Judgement code	The judgement code for describing the patient's condition
Textual data	
Chief complaint	The patient's description of the symptoms
Judgment description	The record that describes the patient's symptoms written by the nursing staff

[a]ED: emergency department.

[b]IC: integrated circuit.

Data Augmentation

After analyzing our prospective data set, we observed an imbalanced data distribution. As machine learning algorithms tend to increase accuracy by reducing errors, most of them are biased toward the majority class and tend to ignore the minority class. For instance, 758 out of 901 (84.1%) ED patients were discharged from the hospital in our prospective data set, and the system could achieve 85% accuracy if it kept on predicting discharge. However, we did not want the system to only indicate discharge. Therefore, to avoid the above situation, we used the "synthetic minority oversampling technique" (SMOTE) to generate some synthesized data to ensure that the system could learn the different patterns between each class. In our study, the iteration of the SMOTE algorithm started by selecting 1 minority sample and finding its top 5 nearest neighbors. These 5 neighbors were chosen to generate new synthesized data by the interpolation method. Finally, the iteration was repeated several times until we obtained the minority class where the number was the same as that of the majority class. However, as the synthesized data may be too diverse, some of the data can have negative influences on the model. Therefore, we used the Tomek Links algorithm to remove some ambiguous data that may hurt model performance by pairing samples and removing the pairs with different labels. An example of the augmentation process is shown in Multimedia Appendix 3. In the original data set, we can observe that only 143 patients are admitted. After applying the SMOTE algorithm on our data set, the number of admitted patients increases to 758. We then use the Tomek Links algorithm to remove some samples that are regarded as

ambiguous samples by the algorithm. Finally, in this example, a total of 1294 patients are included in our new augmented prospective data set.

As for text data, since the SMOTE algorithm cannot generate text, we set up a mapping relation to add the text feature for each synthesized sample. First, we created a number of lists, each of which stores the chief complaints from data samples sharing the same class label. After these lists and the synthesized data were ready, for each synthesized sample, we randomly selected 1 chief complaint from the list according to its label and added it as a text feature of the synthesized sample.

Pretraining of the Vital Sign Encoder

The TabNet architecture is composed of feature transformers and attentive transformers. In TabNet's design, the mask from the attentive transformer can select the most vital feature from several features, eliminating noise caused by irrelevant features. Furthermore, the mask can be calculated to provide some interpretable information about the feature's importance. Therefore, considering the objective of this study, our work takes advantage of the encoder-decoder architecture of TabNet, which is inspired by Arik [18], and we adopted this architecture to construct our vital sign encoder (Figure 4).

Before training on the prospective data set, the vital sign encoder was pretrained on retrospective data by unsupervised learning to learn some basic information about such structural data. Structural features of demographics, triage information, and vital sign information (Table 1) were used in this step.

Figure 5 shows the process used for pretraining our vital sign encoder. In triage level prediction and length of stay prediction, since we did not have a triage golden label and length of stay label for pretraining the vital sign encoder, we used only unsupervised learning. On the other hand, both unsupervised learning and supervised learning were used for hospitalization prediction. The reason why we used the unsupervised learning algorithm is that the model can discover hidden data patterns without human intervention by analyzing and clustering the unlabeled information.

Figure 4. Vital sign encoder architecture (adapted from TabNet). FC: fully connected networks; ReLU: rectified linear unit.

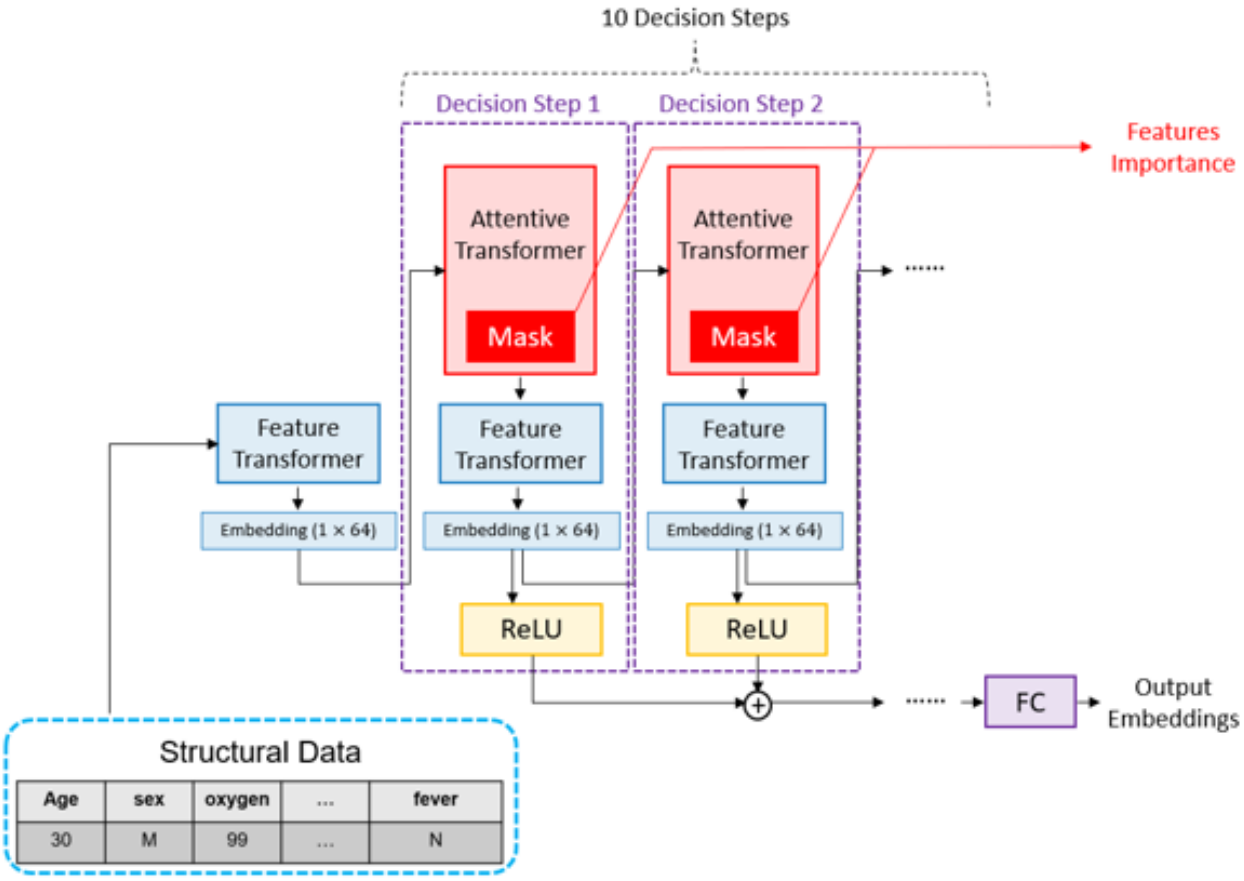
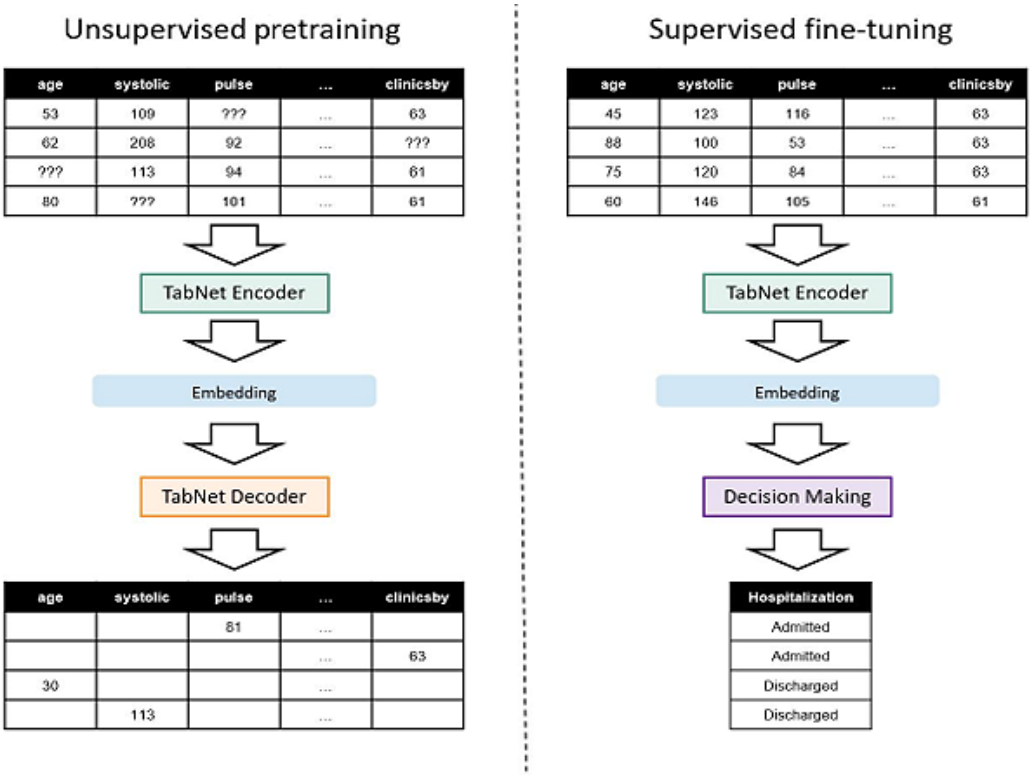


Figure 5. The flow of pretraining the vital sign encoder.



In our study, we used the encoder-decoder architecture. We masked some of the input features in our data and then reconstructed the masked features. The reconstruction loss during unsupervised learning is described as follows:

where B denotes the batch size, D denotes the dimension of features (number of features), S is a binary mask sparse matrix with size B×D for masking some of the features, and $f_{B\times D}$ and  are matrices of features with size B×D representing the ground truth feature values and the predicted outputs, respectively [18].

The purpose of Equation 1 is to calculate the distance between the ground truth feature and the predicted feature. In each iteration, the binary mask $S_{b,j}$ is sampled independently from Bernoulli distribution, and the mask can only have a value of 0 or 1. During this process involving the masked value and its reconstruction, models are believed to learn implicit relationships between features.

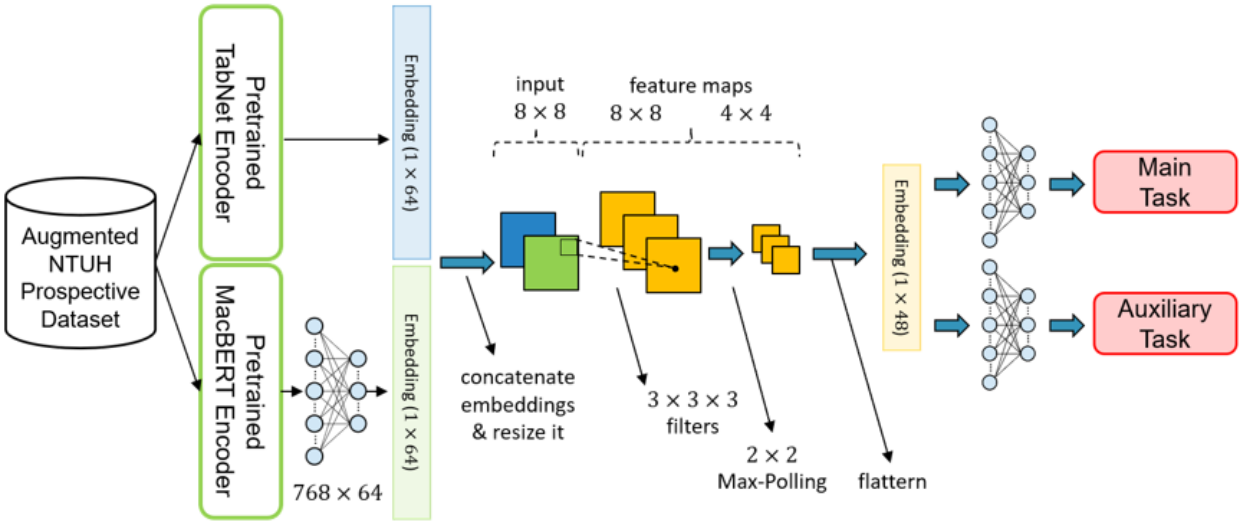
Pretraining of the MacBERT Encoder

BERT is a well-known language model that can be used to transform a word into a representation and understand the meaning behind the sentence. In addition, it performs consistently better than other language models (eg, ELMo [12] and GPT [14]) and also performs well in many different tasks. However, although the BERT model can be easily fine-tuned with an additional output layer to achieve outstanding performance, the pretraining process of the model is designed for general purposes. In this study, to better understand our medical data, we repretrained MacBERT (Chinese version of BERT) by applying MLM again. We extracted the text information from the NTUH prospective data set and then used the information to accomplish further training of MLM. All the settings of the training process followed the original set in MacBERT. By further training with MLM, the fine-tuned MacBERT could enrich its knowledge in specific domains.

Overall Model Architecture

The typical model architecture of each subsystem is shown in Figure 6. After the pretrained encoders are ready, the encoder weights are copied to the fine-tuning stage encoders. The typical model architecture can be divided into 4 main parts: input, encoders, classifiers, and output. First, in the input part, there are 2 data types, namely, structural data and free-text data. Since the prospective data set has only a limited amount of data, we sent it to the augmentation algorithm to obtain synthesized data and added them to the original data set. Second, the structural data and the free-text data are sent to the pretrained TabNet encoder and the pretrained MacBERT encoder, respectively. Afterward, to obtain a comprehensive representation of the data, 2 embeddings coming out from the pretrained TabNet encoder and pretrained MacBERT encoder are concatenated together. Third, the concatenated embeddings are passed through classifiers for output prediction.

Figure 6. Typical model architecture in the fine-tuning stage. MacBERT: Chinese version of bidirectional encoder representations from transformers; NTUH: National Taiwan University Hospital.



Input

We used the augmented NTUH prospective data set in the fine-tuning stage. The data set contains 2 data types. The first is structural data, including patient demographics, triage information, and vital sign information. The second is free-text data, including patient chief complaints, nursing staff judgment descriptions, and transformed information from the structural data (Multimedia Appendix 4). However, since MacBERT is a Chinese BERT model, which is trained on simplified Chinese, we translated our text data from traditional Chinese to simplified Chinese to achieve better performance.

Encoders

As shown in Figure 6, since there were 2 types of data to be processed, we used the TabNet encoder and MacBERT encoder to extract feature information from structural data and free-text data, respectively. We then transformed these information pieces into high-dimensional embeddings for further training.

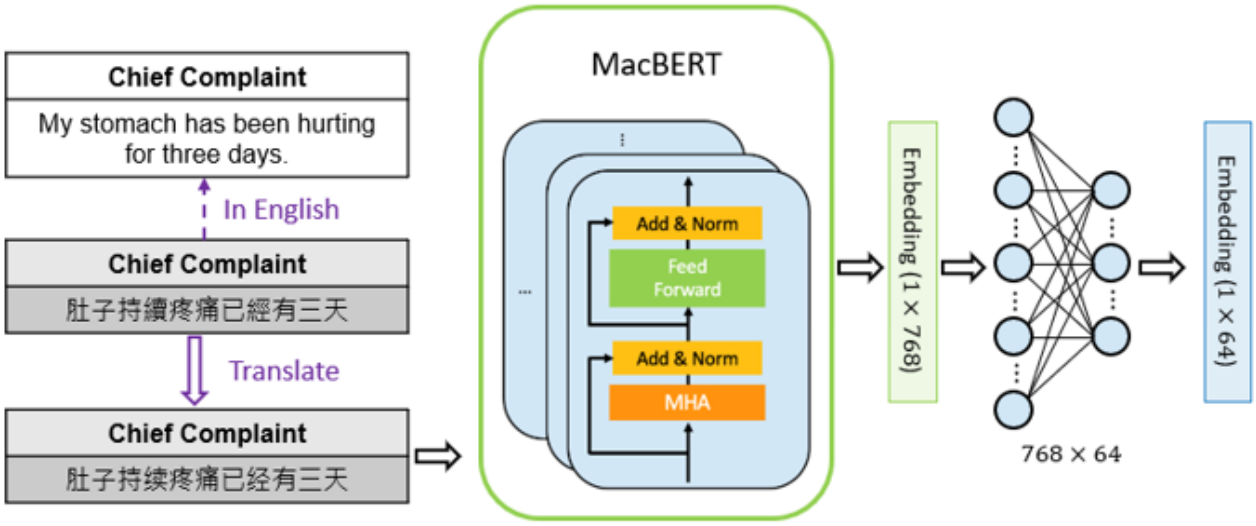
Pretrained Vital Sign Encoder

We used the pretrained TabNet encoder as our vital sign encoder. In the pretraining stage, we obtained some basic information of these medical data from the NTUH retrospective data set. As a result, to achieve better starting, the pretrained weights were directly deployed into our vital sign encoder. We stacked up 10 decision steps to build our vital sign encoder, and the dimensions of both the input and output were set to 64. A 1×64 vector was the final context vector.

Pretrained Language Model Encoder

As chief complaints are manually recorded by nurses and most of them are written in traditional Chinese, it is better to find a language model that has been trained on a Chinese corpus and can handle Chinese text well. MacBERT is an improved BERT model with novel MLM as a correction pretraining task, which mitigates the discrepancy between pretraining and fine-tuning. Moreover, it has been trained on simplified Chinese corpora, which is more suitable for our work. As a result, we decided to adopt MacBERT from Hugging Face as the chief complaint text encoder in our proposed model, instead of the original BERT model. On the other hand, we observed that the text in our data set might contain different languages, including English and Chinese. Therefore, to make MacBERT applicable to our case, we translated the text into a uniform language, namely, simplified Chinese, before sending it into MacBERT. However, since we wanted the contributions from the vital sign encoder and the MacBERT encoder to be comparable, a fully connected layer was placed after the output vector from MacBERT to decrease the vector dimension from 1×768 to 1×64. The entire process explaining how we handled the text data is shown in Figure 7.

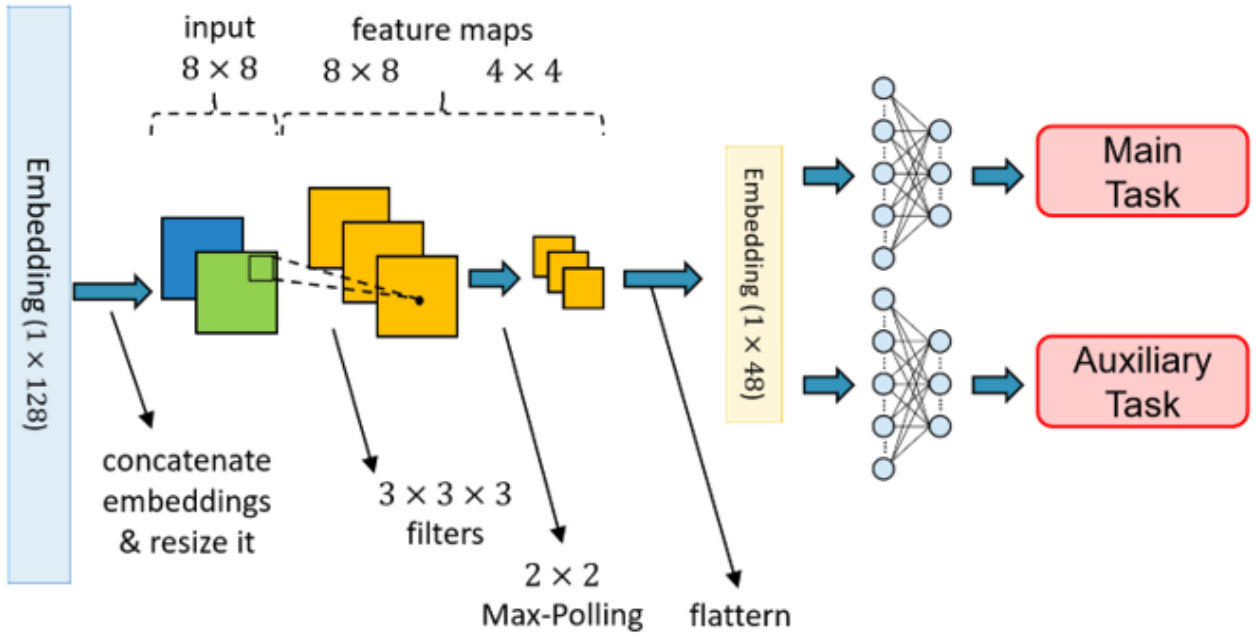
Figure 7. The entire process of handling text data. MacBERT: Chinese version of bidirectional encoder representations from transformers.



Classifiers

All the inputs were encoded into high-dimensional embeddings by the encoders mentioned in the previous stage. It is believed that both embeddings have different facets of information; therefore, instead of adding these vectors together, we concatenated these 2 vectors to obtain richer patient information before sending them into the classifiers. Moreover, in our study, we adopted the multi-task learning architecture to learn shared representation and avoid overfitting problems. As a result, there were 2 classifiers for predicting different targets, where each classifier had a 1-layer convolutional neural network and a 2-layer multi-layer perceptron. The details of the process are shown in Figure 8.

Figure 8. Components of the classifiers.



Output

In contrast to most single-output machine learning methods, our proposed model has a multi-task model architecture. Multi-task learning is a type of machine learning method by which the multi-output outcome can be learned simultaneously in a shared model. In addition to the data efficiency advantages, such an approach can reduce overfitting by leveraging auxiliary information and allowing fast learning. Since target prediction loss will update the encoders, the encoders can avoid being overfitted and learn more general knowledge. As there were 3 medical outcomes in our system, we designed 3 models with slight differences to handle different tasks. The details of these 3 models are shown in Figures 9 to 11.

Figure 9. The model architecture of triage level prediction. FL: focal loss; MacBERT: Chinese version of bidirectional encoder representations from transformers; NTUH: National Taiwan University Hospital.

$$Loss_{total} = FL_{triage} + 0.5 \times FL_{hos} + 10^{-3} \times Loss_{sparse}$$

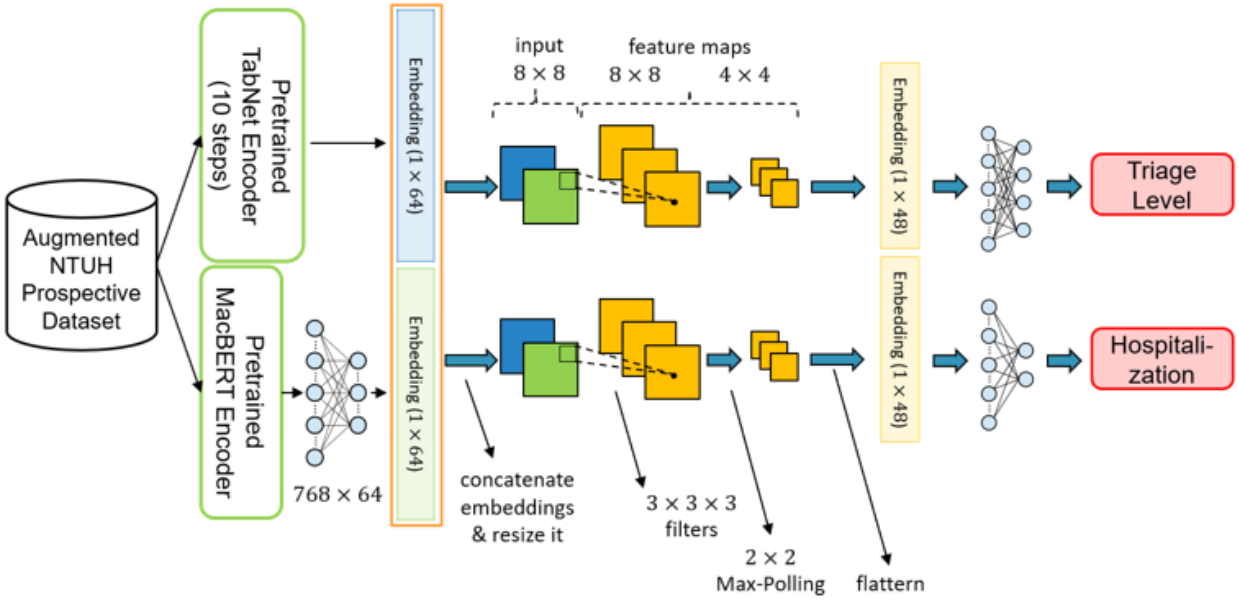


Figure 10. The model architecture of hospitalization prediction. FL: focal loss; MacBERT: Chinese version of bidirectional encoder representations from transformers; NTUH: National Taiwan University Hospital.

$$Loss_{total} = FL_{hos} + 0.3 \times FL_{triage} + 10^{-3} \times Loss_{sparse}$$

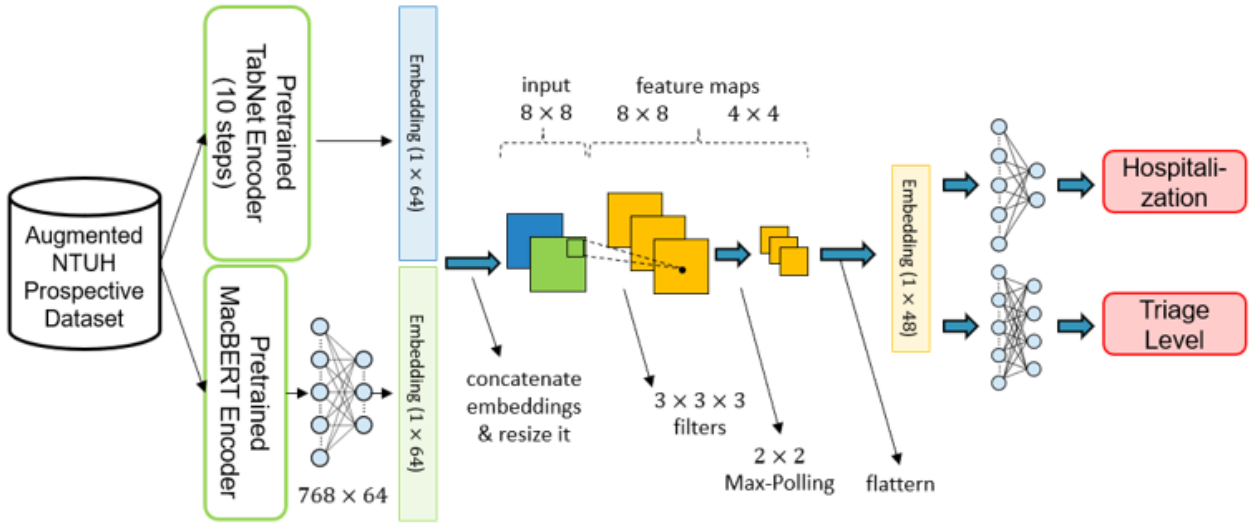
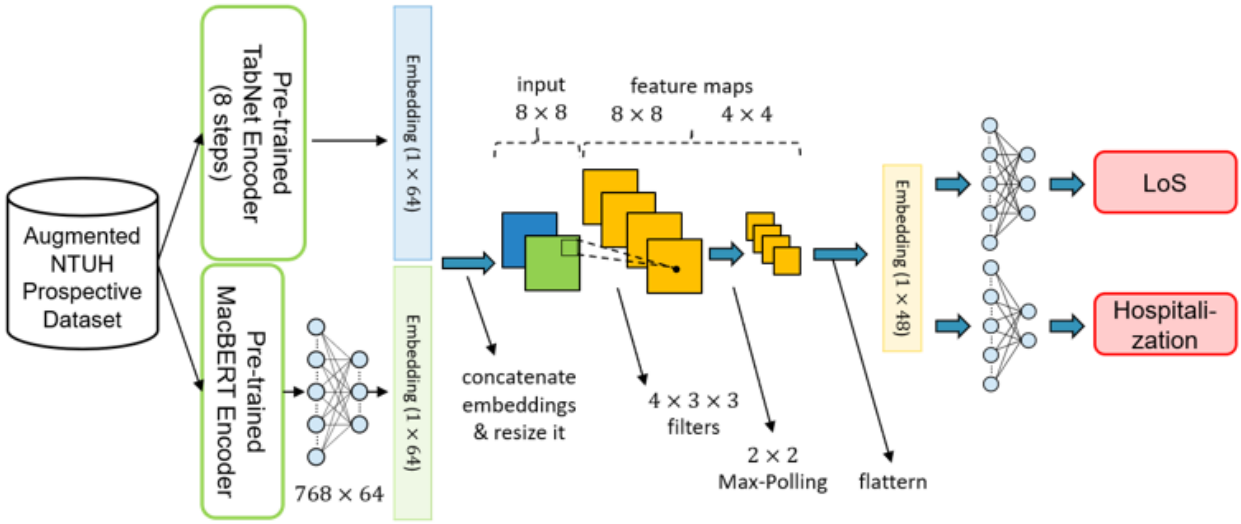


Figure 11. The model architecture of length of stay (LoS) prediction. FL: focal loss; MacBERT: Chinese version of bidirectional encoder representations from transformers; NTUH: National Taiwan University Hospital.

$$Loss_{total} = FL_{LoS} + 0.3 \times FL_{hos} + 10^{-3} \times Loss_{sparse}$$



Loss Function

Total loss combines focal loss and sparse entropy loss as follows:

where λ_1 is a hyperparameter for determining the learning direction of the model via controlling the balance between the main task and related task, and λ_{sparse} is a hyperparameter for controlling the sparsity of the TabNet encoder, where a greater parameter is associated with a greater effect of the tabular data on the entire model, and the TabNet encoder tends to select 1 feature in 1 decision step.

In order to assess the performance of the model, the focal loss function was utilized by comparing the ground truth label with the probability distributions over network predictions, which has been shown as follows:

where $\hat{y}$ is the model prediction, y is the ground truth value, superscript i refers to sample i, y_k is 0 or 1 (indicating whether a class label is the correct classification among K classes),  denotes the confidence score of class k, and γ is a hyperparameter that is set to 2 in our study.

TabNet uses sparse entropy loss (first proposed in [23]) to provide a favorable inductive bias for data sets where most features are redundant. The sparse entropy loss can not only help the model to select salient features from all attributes of the sample, but also fasten the training process. The equation is as follows:

where N_{steps} denotes how many decision steps are stacked up in the model, B is the batch size, D is the total number of features, M represents the mask, $M_{b,j}$ [i] refers to the mask at the i^{th} step with batch sample b and feature j, and ε is a small number to maintain numerical stability.

Results

Experimental Setup

A series of experiments were conducted to validate the effectiveness of our design. The details of our system environment are presented below. We conducted our experiments on the Ubuntu 20.04 operating system with PyTorch 1.7.1 and Python 3.9.7, and all training procedures were performed on a computer with a Nvidia RTX 3090 graphics card, an Intel Core i7-1070K processor, and 32 GB of RAM.

Training Settings

The Adam optimizer with an initial learning rate of 0.01 was used in our experiments, and it was adjusted by the "ReduceLROnPlateau" scheduler with the patient value set as 15. Meanwhile, if the loss did not improve for 50 epochs, an early stop action was taken.

All experiments were carefully conducted in the following steps: (1) The data set was divided into 3 parts (training set, validation set, and testing set in the ratio of 8:1:1); (2) The training set was used to generate synthesized data to make up the gap between classes, and the synthesized data were added into the original training data set; (3) Our design was evaluated by taking the average test performance for 10 trials, as the division of the data set might have varied effects on the experiment results.

Evaluation Metrics

Since our data set was obviously imbalanced, the accuracy performance cannot represent the effectiveness of our system. As a result, in our experiment, the evaluation metrics included

precision, recall, and F1-score. Precision measures the rate of ground truth classes that are predicted correctly. Recall measures the portion of each class of our prediction that is actually that class. Finally, F1-score represents the harmonic mean between precision and recall. Their formulas are as follows:

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

Data Characteristics

Our study included 2 data sets. One data set was the NTUH retrospective data set, which contains a collection of the past EHRs of 268,716 visits from 2009 to 2015, and the other data set was the NTUH prospective data set, which contains data collected with patient consent in the NTUH ED from May 26, 2020, to February 21, 2022, and includes 901 ED patient records after removal of unreasonable and missing data. Table 2 summarizes the data characteristics of vital sign information in these 2 data sets. Despite similar average values across all fields in the 2 data sets, on performing statistical tests using *P*-values, we found that there was a significant difference between the 2 data sets. However, we believe that using data with the same data collection background but different distributions can still effectively improve the robustness and generalization ability of the model. By pretraining on diverse data, the model can learn more general representations, leading to improvements in the final predictions.

On the other hand, the distributions for different tasks are shown in Multimedia Appendix 5. It is worth mentioning that the distribution gap of the triage level between the retrospective data set and prospective data set was greater than the distribution gaps for hospitalization and length of stay. This is because hospitalization and length of stay are based on facts, and in contrast to the triage level in the retrospective data set, the triage level in the prospective data set comes from physician diagnosis. As it is believed that the doctor's triage level can assign patient acuity more accurately, we used it as our golden label for predicting the triage level. Another reason for the distribution gap could be the difficulty in collecting data from more severe patients.

Table 2. Patient characteristics in the National Taiwan University Hospital retrospective and prospective data sets.

Variable	NTUH[a] retrospective data set	NTUH prospective data set
Age (years), mean (SD)	49.1 (19.98)	52.4 (18.98)
Sex, n (%)		
Female	141,783 (52.8)	450 (50.1)
Male	126,933 (47.2)	450 (49.9)
Arrival time, n (%)		
7 AM to 3 PM	10,2256 (42.8)	518 (57.4)
3 PM to 11 PM	11,4970 (38.0)	289 (32.1)
11 PM to 7 AM	5,1490 (19.2)	94 (10.5)
Systolic blood pressure (mmHg), mean (SD)	136.3 (26.79)	132.4 (24.78)
Diastolic blood pressure (mmHg), mean (SD)	80.8 (15.22)	79.8 (13.91)
Pulse (beats/min), mean (SD)	88.8 (18.74)	89.5 (18.74)
Oxygen saturation (%), mean (SD)	97.0 (3.09)	97.7 (1.69)
Respiration (breaths/min), mean (SD)	18.2 (2.16)	18.8 (2.04)
Body temperature (°C), mean (SD)	37.0 (0.82)	36.7 (0.65)
Pain index (scale), n		
0	134,292	357
1-3	9,554	368
4-6	60,526	140
7-10	64,344	36

[a]NTUH: National Taiwan University Hospital.

Experimental Results

We compared our model's performance regarding triage level, hospitalization, and length of stay against the performance of other machine learning methods. As the data of only 901 ED visits were finally included in our study, it was a challenge to obtain a robust model with great capability to identify critical patients.

Unlike other work on triage level prediction, since we endeavored to fix the bias of traditional rule-based system triage, such as the ESI and TTAS, we used the diagnosis results provided by the physician as our golden label. As shown in Table 3, it is worth noting that our triage model achieved a nearly 30% improvement in 4 metrics, including accuracy, precision, recall, and F1-score, when compared to the results obtained from other models. These outstanding results show the promising potential of our proposed model.

As shown in Table 4, we can observe that our hospitalization model achieved the highest performance in 3 metrics, including precision, recall, and F1-score. Although the support vector machine (SVM) model achieved an accuracy of 91.2%, it may tend to predict the majority (discharge) owing to the low precision and recall values. From the previous discussion, it can be seen that our model is the most discriminative model.

Additionally, our proposed model outperformed other models. Although the study design and data set in our study are different from those in other studies, it is worth indicating that with the help of retrospective data pretraining, the model can learn more than with only the use of prospective data. Our proposed model achieved promising results, with 3%-6% improvement in accuracy (Table 5).

As shown in Table 6, although most of the models achieved an accuracy of higher than 70%, their performances on other metrics revealed that these models tend to predict the majority class. Nevertheless, except for accuracy, our length of stay model outperformed other machine learning methods in the other 3 metrics, indicating the capability of our length of stay model for discrimination.

Table 3. Performance comparison between our model and other machine learning methods in the “triage level” task.

Method	Accuracy	Precision	Recall	F1-score
TabNet [18]	0.425	0.436	0.410	0.423
NODE[a] [19]	0.472	0.324	0.328	0.324
Random forest [24]	0.354	0.506	0.300	0.376
XGBoost[b] [15]	0.351	0.394	0.308	0.345
SVM[c] [25]	0.340	0.581	0.268	0.367
Our model	0.633[d]	0.686[d]	0.633[d]	0.658[d]

[a]NODE: neural oblivious decision ensembles.

[b]XGBoost: extreme gradient boosting.

[c]SVM: support vector machine.

[d]Highest value.

Table 4. Performance comparison between our model and other machine learning methods in the “hospitalization” task.

Methods	Accuracy	Precision	Recall	F1-score
TabNet [18]	0.791	0.701	0.702	0.701
NODE[a] [19]	0.752	0.622	0.689	0.653
Random forest [24]	0.821	0.765	0.674	0.717
XGBoost[b] [15]	0.829	0.651	0.679	0.655
SVM[c] [25]	0.912[d]	0.456	0.500	0.477
Our model	0.822	0.811[d]	0.823[d]	0.817[d]

[a]NODE: neural oblivious decision ensembles.

[b]XGBoost: extreme gradient boosting.

[c]SVM: support vector machine.

[d]Highest value.

Table 5. Performance comparison between our model and the models in other related studies in the "hospitalization" task.

Study	Data set	Study type	Accuracy	Precision	Recall	F1-score
Study by Raita et al [26]	NHAMCS[a]	Retrospective	—[b]	—	0.750	—
Study by Yao et al [27]	NHAMCS	Retrospective	0.775	0.820[c]	0.790	0.804
Study by Leung et al [28]	NTUH[d]	Prospective	0.805	0.806	0.790	0.798
Our study	NTUH	Prospective	0.822[c]	0.811	0.823[c]	0.817[c]

[a]NHAMCS: National Hospital Ambulatory Medical Care Survey.

[b]Not reported.

[c]Highest value.

[d]NTUH: National Taiwan University Hospital.

Table 6. Performance comparison between our model and other machine learning methods in the "length of stay" task.

Methods	Accuracy	Precision	Recall	F1-score
TabNet [18]	0.683	0.654	0.665	0.659
NODE[a] [19]	0.721	0.616	0.589	0.602
Random forest [24]	0.754	0.606	0.444	0.512
XGBoost[b] [15]	0.744	0.523	0.446	0.481
SVM[c] [25]	0.791[d]	0.263	0.333	0.294
Our model	0.713	0.786[d]	0.713[d]	0.747[d]

[a]NODE: neural oblivious decision ensembles.

[b]XGBoost: extreme gradient boosting.

[c]SVM: support vector machine.

[d]Highest value.

Ablation Studies

Effectiveness of Multimodality

Experiments were conducted to demonstrate the superior performance of our proposed model. Since our model comprised the TabNet encoder and the language model encoder, we designed an experiment to show that the performance of a model leveraging both vital sign information and text information is better than that of a model using only 1 information modality. Table 7 shows that the proposed model achieved the best performance when both modalities were used. The results suggest that both structural and text data contribute to model prediction. The greater performance of the model using only tabular data than that using only text data could be attributed to the advantage of pretraining, as the vital sign encoder was pretrained with a large volume of retrospective data.

Table 7. The effectiveness of different modalities in the "triage level" task.

Methods	Accuracy	Precision	Recall	F1-score
Only tabular data	0.575	0.613	0.568	0.589
Only text data	0.439	0.119	0.250	0.162
Our method (tabular data + text data)	0.633[a]	0.686[a]	0.633[a]	0.658[a]

[a]Highest value.

Effectiveness of Multitask Training and Data Augmentation

Multitask learning experiments confirmed that the approach does offer advantages like improving data efficiency, reducing overfitting through shared representations, and allowing fast learning by leveraging auxiliary information. However, in order to obtain a more robust feature extractor, in a general setting, the targets in the multitask learning model should be related. As a result, in the experiments, we selected triage level prediction and hospitalization as our 2 outputs. It is believed that a patient assigned to level 1 or 2 should have a higher probability of admission to the hospital after being discharged from the ED. Moreover, since data distribution in triage labels is unbalanced, we attempted to narrow the distribution gap by using the method of data augmentation. Table 8 shows that both multitask learning and augmentation contributed to better performance.

Table 8. The effectiveness of different architectures in the "triage level" task.

Methods	Accuracy	Precision	Recall	F1-score
Multitask	0.500	0.369	0.500	0.425
Single task + augmentation	0.583	0.600	0.582	0.591
Single task	0.458	0.506	0.455	0.479
Our method (multitask + augmentation[a])	0.633[b]	0.686[b]	0.633[b]	0.658[b]

[a]The method of data augmentation used in our proposed model is described in the "Data Augmentation" subsection.

[b]Highest value.

Effectiveness of Different Language Models

Experiments were conducted to evaluate the performance between different language models (Table 9). In our original data set, the chief complaint was written in traditional Chinese. However, no language model has been trained on traditional Chinese. Hence, to solve this problem, we first translated the text features into different languages before sending them to the respective language models. The results showed that the model using MacBERT as the language encoder was better than models using other approaches.

Table 9. The effectiveness of different language models in the "triage level" task.

Methods	Data language	Accuracy	Precision	Recall	F1-score
Multilingual BERT[a]	Simplified Chinese	0.500	0.369	0.500	0.425
Multilingual BERT	English	0.583	0.600	0.582	0.591
BERT	English	0.458	0.506	0.455	0.479
Our method (MacBERT[b])	Simplified Chinese	0.633[c]	0.686[c]	0.633[c]	0.658[c]

[a]BERT: bidirectional encoder representations from transformers.

[b]MacBERT: Chinese version of BERT.

[c]Highest value.

Effectiveness of Different Fusion Methods

Experiments were conducted to demonstrate the superior performance of our proposed model. As our model directly concatenated the decreased embedding from the language model and the embedding from the vital sign encoder, we designed an experiment to show that it is necessary to make contributions for the text data and structural data to be comparable, and direct concatenation fusion can preserve more information than addition fusion. In Table 10, the first experiment involves the model adding 2 embeddings (text and vital sign embeddings) together with a learnable scale value to balance the gap between the text and vital sign embeddings, and the second experiment involves directly using the embedding from the language model instead of passing another fully connected network to decrease its dimension. The results suggest that making 2 embeddings to be comparable and using a direct concatenation fusion method can contribute to better performance.

Table 10. The effectiveness of different fusion methods in the "triage level" task.

Methods	Accuracy	Precision	Recall	F1-score
Experiment 1 (addition fusion)	0.548	0.580	0.547	0.563
Experiment 2 (no concatenation fusion)	0.583	0.634	0.583	0.607
Our method	0.633[a]	0.686[a]	0.633[a]	0.658[a]

[a]Highest value.

Interpretability

Although machine learning models can provide remarkably good prediction results, models need to provide explanations of the results that humans can understand easily. In our proposed model, for structural features, the attentive transformer from TabNet generated the mask to mask out different features in each decision step and observed how these features affect the model performance. As a final step, the attentive transformer calculated the importance of features by adding up the mask values of each step. On the other hand, BertViz [29] is an interactive tool that can visualize attention in transformer language models such as BERT. By acquiring attention scores from transformer layers in language models, BertViz can point out important words that contribute to the predicted result.

Multimedia Appendix 6 provides an inference example from the field test, and Multimedia Appendix 7 provides the prediction results of the inference sample for hospitalization. In this example, the patient shows acute change during the triage process, extremely high systolic and diastolic blood pressure, and an unusual Glasgow Coma Scale (GCS) score. As shown

in Multimedia Appendix 7 our system recommended admission of the inferenced patient, and the patient was actually admitted to the hospital. Our system not only successfully provided the correct suggestion to the nursing staff, but also indicated that acute change, systolic blood pressure, diastolic blood pressure, GCS-E, and GCS-M have important effects on the prediction result. As for text analysis, we used the concept from BertViz to extract attention scores for each token from the language model and visualize these attention scores. Although the language model had a hierarchy of linguistic signals from phrase to semantic features, it is believed that the deeper layer of the language model holds more information of the whole sentence [30]. Hence, we extracted the attention score from the ninth layer of the language model for further visualization (Multimedia Appendix 8).

System Application

Triage aims to prioritize patients in the ED and ration care toward those patients who need immediate care. However, recently, owing to the rising number of elderly patients and the high volume of low-acuity ED visits under waiting, patients tend to wait for very long to see the physician. This situation can cause several severe clinical outcomes such as increased mortality rates.

With the advancement in technology and popular application of computers nowadays, we wonder whether machine learning methods can help to mitigate the overcrowding problem in the ED. Therefore, we developed a triage system based on our proposed model and adopted it in the NTUH ED to provide stable and reasonable clinical AI suggestions to nursing staff. For application in the real world, we should take the running time of the system into account. The entire running time of each part is shown in Multimedia Appendix 9. The system takes no more than 10 seconds to make clinical predictions.

Before the system is officially launched, we planned a field test to ensure that the system can achieve promising performance in the real world. Finally, we included almost 6500 ED patients in our analysis from September 30, 2022, to December 30, 2022. The distributions of hospitalization and length of stay between these patients were quite different as compared to the NTUH prospective data set (Multimedia Appendix 10 and Multimedia Appendix 11). Especially for length of stay, patients who stayed in the ED for over 24 hours were much less in this data set than in both NTUH data sets (Multimedia Appendix 11). Moreover, since our golden triage level depended on the physician's diagnosis, it was challenging to label all patients in the field test; however, we evaluated our system in another way, which will be discussed later. The distribution gap between both NTUH data sets and the field test is presented in Multimedia Appendix 12.

As shown in Multimedia Appendix 13 and Multimedia Appendix 14, there was a slight performance gap between the experiments on the earlier mentioned data sets and the real-world data. However, from the results of the confusion matrix, it can be seen that in the case of "patients actually discharged," 2085 out of 2539 (82.1%) discharged patients were accurately predicted and were recommended to be discharged by the system. On the other hand, in the case of "patients actually admitted," 194 out of 316 (61.4%) patients were accurately predicted and were recommended to stay in the hospital.

As mentioned previously, for length of stay, there was a large distribution gap between our field test data set and the NTUH prospective data set. Multimedia Appendix 15 and Multimedia Appendix 16 show that the system cannot perform as good as it does in local experiments. However, from the results of the confusion matrix, we can observe that the system has a better capability of discriminating patients who stay for less than 6 hours, and the system tends to underestimate patients who stay in the ED for 6 to 24 hours.

Finally, Multimedia Appendix 17 and Multimedia Appendix 18 show that although the newly collected data did not have the golden triage level labels provided by the doctors, the distribution of the triage level indicated that the model predicted a fairly even distribution, while the system triage still mainly predicted level 3.

Discussion

Limitations

Although our proposed model showed good preliminary results compared to the results of other machine learning methods, it still has a long way to go. For instance, despite our model's ability to incorporate various language models, it may not perform well for languages where specific language models are not available in the training data set. Second, as we need to translate the text into a uniform language initially and the sentence in the data is not always complete, a better translator and some postprocessing techniques are needed to alleviate the problems. Additionally, as retrospective data lack a label in triage level prediction, expansion of the data set for training the model should help the model to learn a wider range of patterns and should enhance model performance. Moreover, since our proposed model can allow efficient learning of image or text encoders in the presence of multimodality along with tabular data, further work can add images or speech information into our model to help it achieve better performance.

Conclusion

Emergency services are an essential aspect of the health care system in hospitals, and the demand for these services has increased exponentially in recent years. Although Taiwan has established a standard process of assigning patients to different emergency levels, there is insufficient capacity to ensure precise assignment. Most patients are over-triaged or under-triaged, which can waste limited medical resources or have severe consequences such as patient mortality.

In this study, we aimed to design a deep learning prediction system that can prioritize patients and assign patients to appropriate triage levels. To obtain rich information from patients, our proposed model not only uses vital sign information, but also leverages text information.

Our system included a well-pretrained vital sign encoder and a repretrained MacBERT encoder. Additionally, by using the multitask learning and data augmentation method, we successfully obtained promising results for triage level

prediction, hospitalization prediction, and length of stay prediction. For triage level prediction, there were nearly 30% improvements in 4 metrics compared with other machine learning methods, including accuracy, precision, recall, and F1-score. Different modalities and model architectures have also been studied for ablation effectiveness. Moreover, our proposed model also provides clinicians with interpretability to understand the reasons behind the model predictions.

In conclusion, our system improved the prediction of 3 different medical outcomes when compared with other machine learning methods. With the pretrained vital sign encoder and repretrained MLM MacBERT encoder, our multimodality model can provide a deeper insight into the characteristics of EHRs. Additionally, by providing interpretability, we believe that the proposed system can assist nursing staff and physicians in taking appropriate medical decisions.

Acknowledgments

This research was supported by the Joint Research Center for AI Technology and All Vista Healthcare under the Ministry of Science and Technology of Taiwan (grants 111-2223-E-002-008 and 111-2634-E-002-021), and by the Center for Artificial Intelligence and Advanced Robotics, National Taiwan University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Comparison between related studies and our proposed work.
[PNG File , 338 KB - medinform_v12i1e48862_app1.png]

Multimedia Appendix 2
An example of uncleaned data.
[PNG File , 54 KB - medinform_v12i1e48862_app2.png]

Multimedia Appendix 3
An example of the number change in the data using the synthetic minority oversampling technique (SMOTE) algorithm and Tomek Links algorithm.
[PNG File , 87 KB - medinform_v12i1e48862_app3.png]

Multimedia Appendix 4
The data construction of the input (left is the structural data; right is the free-text data).
[PNG File , 603 KB - medinform_v12i1e48862_app4.png]

Multimedia Appendix 5
Distribution of each task.
[PNG File , 286 KB - medinform_v12i1e48862_app5.png]

Multimedia Appendix 6
An inference example from the field test.
[PNG File , 279 KB - medinform_v12i1e48862_app6.png]

Multimedia Appendix 7
Prediction result and feature importance of the inferenced patient for hospitalization from the field test (structural data).
[PNG File , 185 KB - medinform_v12i1e48862_app7.png]

Multimedia Appendix 8
Text visualization of the inference patient for hospitalization from the field test.
[PNG File , 520 KB - medinform_v12i1e48862_app8.png]

Multimedia Appendix 9
The running time of each part in the system.
[PNG File , 85 KB - medinform_v12i1e48862_app9.png]

Multimedia Appendix 10
The distribution gap between both National Taiwan University Hospital data sets and the field test for hospitalization.
[PNG File , 42 KB - medinform_v12i1e48862_app10.png]

Multimedia Appendix 11
The distribution gap between both National Taiwan University Hospital data sets and the field test for length of stay.
[PNG File , 58 KB - medinform_v12i1e48862_app11.png]

Multimedia Appendix 12
The distribution gap between both National Taiwan University Hospital data sets and the field test.
[PNG File , 44 KB - medinform_v12i1e48862_app12.png]

Multimedia Appendix 13
The performance of hospitalization in the field test.
[PNG File , 28 KB - medinform_v12i1e48862_app13.png]

Multimedia Appendix 14
The performance (truth) of hospitalization in the field test.
[PNG File , 24 KB - medinform_v12i1e48862_app14.png]

Multimedia Appendix 15
The performance of length of stay in the field test.
[PNG File , 29 KB - medinform_v12i1e48862_app15.png]

Multimedia Appendix 16
The performance (truth) of length of stay in the field test.
[PNG File , 45 KB - medinform_v12i1e48862_app16.png]

Multimedia Appendix 17
The prediction of triage in the field test.
[PNG File , 51 KB - medinform_v12i1e48862_app17.png]

Multimedia Appendix 18
Graph showing the prediction of triage in the field test.
[PNG File , 31 KB - medinform_v12i1e48862_app18.png]

References

1. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS One 2018 Jul 20;13(7):e0201016 [FREE Full text] [doi: 10.1371/journal.pone.0201016] [Medline: 30028888]
2. Guttmann A, Schull MJ, Vermeulen MJ, Stukel TA. Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada. BMJ 2011 Jun 01;342(jun01 1):d2983-d2983 [FREE Full text] [doi: 10.1136/bmj.d2983] [Medline: 21632665]
3. Bullard MJ, Unger B, Spence J, Grafstein E, CTAS National Working Group. Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) adult guidelines. CJEM 2008 Mar 21;10(2):136-151. [doi: 10.1017/s1481803500009854] [Medline: 18371252]
4. Ng C, Yen Z, Tsai JC, Chen LC, Lin SJ, Sang YY, TTAS national working group. Validation of the Taiwan triage and acuity scale: a new computerised five-level triage system. Emerg Med J 2011 Dec 12;28(12):1026-1031. [doi: 10.1136/emj.2010.094185] [Medline: 21076055]
5. Tanabe P, Travers D, Gilboy N, Rosenau A, Sierzega G, Rupp V, et al. Refining Emergency Severity Index Triage Criteria. Acad Emergency Med 2005 Jun;12(6):497-501. [doi: 10.1111/j.1553-2712.2005.tb00888.x]
6. Choi SW, Ko T, Hong KJ, Kim KH. Machine Learning-Based Prediction of Korean Triage and Acuity Scale Level in Emergency Department Patients. Healthc Inform Res 2019 Oct;25(4):305-312 [FREE Full text] [doi: 10.4258/hir.2019.25.4.305] [Medline: 31777674]
7. Tanabe P, Gimbel R, Yarnold PR, Kyriacou DN, Adams JG. Reliability and validity of scores on The Emergency Severity Index version 3. Acad Emerg Med 2004 Jan 08;11(1):59-65 [FREE Full text] [doi: 10.1197/j.aem.2003.06.013] [Medline: 14709429]

8. 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 28;7(3):236-242 [FREE Full text] [doi: 10.1111/j.1553-2712.2000.tb01066.x] [Medline: 10730830]
9. Wuerz RC, Travers D, Gilboy N, Eitel DR, Rosenau A, Yazhari R. Implementation and refinement of the emergency severity index. Acad Emerg Med 2001 Feb 28;8(2):170-176 [FREE Full text] [doi: 10.1111/j.1553-2712.2001.tb01283.x] [Medline: 11157294]
10. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha, Qatar. [doi: 10.3115/v1/D14-1162]
11. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv. 2013. URL: https://arxiv.org/abs/1301.3781 [accessed 2024-02-19]
12. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv. 2018. URL: https://arxiv.org/abs/1802.05365 [accessed 2024-02-19]
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2024-02-19]
14. Radford A, Narasimhan K, Salimans T, Sutskever I. mproving Language Understanding by Generative Pre-Training. OpenAI. 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2024-02-19]
15. 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 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: 10.1145/2939672.2939785]
16. Prokhorenkova L, Gusev G, Vorobev A, Dorogush A, Gulin A. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018 Presented at: 32nd International Conference on Neural Information Processing Systems; December 3-8, 2018; Montréal, Canada. [doi: 10.5555/3327757.3327770]
17. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach CA. [doi: 10.5555/3294996.3295074]
18. Arik S, Pfister T. TabNet: Attentive Interpretable Tabular Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2021 Presented at: AAAI Conference on Artificial Intelligence; February 2-9, 2021; Virtual p. 6679-6687. [doi: 10.1609/aaai.v35i8.16826]
19. Popov S, Morozov S, Babenko A. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. arXiv. 2019. URL: https://arxiv.org/abs/1909.06312 [accessed 2024-02-19]
20. Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. arXiv. 2020. URL: https://arxiv.org/abs/2012.06678 [accessed 2024-02-19]
21. Ivanov O, Wolf L, Brecher D, Lewis E, Masek K, Montgomery K, et al. Improving ED Emergency Severity Index Acuity Assignment Using Machine Learning and Clinical Natural Language Processing. J Emerg Nurs 2021 Mar;47(2):265-278.e7 [FREE Full text] [doi: 10.1016/j.jen.2020.11.001] [Medline: 33358394]
22. Liu Y, Gao J, Liu J, Walline JH, Liu X, Zhang T, et al. Development and validation of a practical machine-learning triage algorithm for the detection of patients in need of critical care in the emergency department. Sci Rep 2021 Dec 15;11(1):24044 [FREE Full text] [doi: 10.1038/s41598-021-03104-2] [Medline: 34911945]
23. Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In: Proceedings of the 17th International Conference on Neural Information Processing Systems. 2004 Presented at: 17th International Conference on Neural Information Processing Systems; December 1, 2004; Vancouver, British Columbia, Canada. [doi: 10.5555/2976040.2976107]
24. Breiman L. Random Forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
25. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their Applications 1998;13(4):18-28. [doi: 10.1109/5254.708428]
26. Raita Y, Goto T, Faridi M, Brown D, Camargo C, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019 Feb 22;23(1):64 [FREE Full text] [doi: 10.1186/s13054-019-2351-7] [Medline: 30795786]
27. Yao LH, Leung KC, Hong JH, Tsai CL, Fu LC. A System for Predicting Hospital Admission at Emergency Department Based on Electronic Health Record Using Convolution Neural Network. 2020 Presented at: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC); October 11-14, 2020; Toronto, ON. [doi: 10.1109/SMC42975.2020.9282952]
28. Leung KC, Lin YT, Hong DY, Tsai CL, Huang CH, Fu LC. A Novel Interpretable Deep-Learning-Based System for Triage Prediction in the Emergency Department: A Prospective Study. 2021 Presented at: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC); October 17-20, 2021; Melbourne, Australia.
29. Vig J. BertViz: A tool for visualizing multihead self-attention in the BERT model. 2019 Presented at: ICLR 2019 Debugging Machine Learning Models Workshop; May 2019; New Orleans, LA.

30. 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: 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy. [doi: 10.18653/v1/P19-1356]

Abbreviations

AI: artificial intelligence
BERT: bidirectional encoder representations from transformers
CatBoost: category boosting
C-NLP: clinical natural language processing
DT: decision tree
ED: emergency department
EHR: electronic health record
ELMo: embeddings from language models
ESI: Emergency Severity Index
GCS: Glasgow Coma Scale
GPT: generative pretrained transformer
MacBERT: Chinese version of bidirectional encoder representations from transformers
MLM: mask language modeling
NLP: natural language processing
NTUH: National Taiwan University Hospital
SMOTE: synthetic minority oversampling technique
TTAS: Taiwan Triage Acuity Scale
XGBoost: extreme gradient boosting

Edited by A Benis; submitted 10.05.23; peer-reviewed by U Sinha, A Garcia Abejas, D Hu; comments to author 26.09.23; revised version received 20.11.23; accepted 05.01.24; published 01.04.24.

Please cite as:
Lin YT, Deng YX, Tsai CL, Huang CH, Fu LC
Interpretable Deep Learning System for Identifying Critical Patients Through the Prediction of Triage Level, Hospitalization, and Length of Stay: Prospective Study
JMIR Med Inform 2024;12:e48862
URL: https://medinform.jmir.org/2024/1/e48862
doi:10.2196/48862
PMID:38557661

The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation

Yong Nam Gwon[1,*], MD; Jae Heon Kim[1,*], MD, PhD; Hyun Soo Chung[2], MD; Eun Jee Jung[2], MD; Joey Chun[1,3], MD; Serin Lee[1,4], MD; Sung Ryul Shim[5,6], MPH, PhD

1

2

3

4

5

6

*these authors contributed equally

Corresponding Author:
Sung Ryul Shim, MPH, PhD

Abstract

Background: A large language model is a type of artificial intelligence (AI) model that opens up great possibilities for health care practice, research, and education, although scholars have emphasized the need to proactively address the issue of unvalidated and inaccurate information regarding its use. One of the best-known large language models is ChatGPT (OpenAI). It is believed to be of great help to medical research, as it facilitates more efficient data set analysis, code generation, and literature review, allowing researchers to focus on experimental design as well as drug discovery and development.

Objective: This study aims to explore the potential of ChatGPT as a real-time literature search tool for systematic reviews and clinical decision support systems, to enhance their efficiency and accuracy in health care settings.

Methods: The search results of a published systematic review by human experts on the treatment of Peyronie disease were selected as a benchmark, and the literature search formula of the study was applied to ChatGPT and Microsoft Bing AI as a comparison to human researchers. Peyronie disease typically presents with discomfort, curvature, or deformity of the penis in association with palpable plaques and erectile dysfunction. To evaluate the quality of individual studies derived from AI answers, we created a structured rating system based on bibliographic information related to the publications. We classified its answers into 4 grades if the title existed: A, B, C, and F. No grade was given for a fake title or no answer.

Results: From ChatGPT, 7 (0.5%) out of 1287 identified studies were directly relevant, whereas Bing AI resulted in 19 (40%) relevant studies out of 48, compared to the human benchmark of 24 studies. In the qualitative evaluation, ChatGPT had 7 grade A, 18 grade B, 167 grade C, and 211 grade F studies, and Bing AI had 19 grade A and 28 grade C studies.

Conclusions: This is the first study to compare AI and conventional human systematic review methods as a real-time literature collection tool for evidence-based medicine. The results suggest that the use of ChatGPT as a tool for real-time evidence generation is not yet accurate and feasible. Therefore, researchers should be cautious about using such AI. The limitations of this study using the generative pre-trained transformer model are that the search for research topics was not diverse and that it did not prevent the hallucination of generative AI. However, this study will serve as a standard for future studies by providing an index to verify the reliability and consistency of generative AI from a user's point of view. If the reliability and consistency of AI literature search services are verified, then the use of these technologies will help medical research greatly.

(JMIR Med Inform 2024;12:e51187) doi:10.2196/51187

KEYWORDS

artificial intelligence; search engine; systematic review; evidence-based medicine; ChatGPT; language model; education; tool; clinical decision support system; decision support; support; treatment

Introduction

The global artificial intelligence (AI) health care market size was estimated to be at US $15.1 billion in 2022 and is expected to surpass approximately US $187.95 billion by 2030, growing at an annualized rate of 37% during the forecast period from 2022 to 2030 [1]. In particular, innovative applications of medical AI are expected to increase in response to medical demand, which will explode in 2030 [2,3].

A large language model (LLM) is a type of AI model that opens up great possibilities for health care practice, research, and education, although scholars have emphasized the need to proactively address the issue of unvalidated and inaccurate information regarding its use [4,5]. One of the best-known LLMs is ChatGPT (OpenAI). It was launched in November 2022. Similar to other LLMs, ChatGPT is trained on huge text data sets in numerous languages, allowing it to respond to text input with humanlike responses [4]. Developed by the San Francisco–based AI research laboratory OpenAI, ChatGPT is based on a generative pre-trained transformer (GPT) architecture. It is considered an advanced form of a chatbot, an umbrella term for a program that uses a text-based interface to understand and generate responses. The key difference between a chatbot and ChatGPT is that a chatbot is usually programmed with a limited number of responses, whereas ChatGPT can produce personalized responses according to the conversation [4,6].

Sallam's [5] systematic review (SR) sought to identify the benefits and current concerns regarding ChatGPT. That review advises that health care research could benefit from ChatGPT, since it could be used to facilitate more efficient data set analysis, code generation, and literature reviews, thus allowing researchers to concentrate on experimental design as well as drug discovery and development. The author also suggests that ChatGPT could be used to improve research equity and versatility in addition to its ability to improve scientific writing. Health care practice could also benefit from ChatGPT in multiple ways, including enabling improved health literacy and delivery of more personalized medical care, improved documentation, workflow streamlining, and cost savings. Health care education could also use ChatGPT to provide more personalized learning with a particular focus on problem-solving and critical thinking skills [5]. However, the same review also lays out the current concerns, including copyright issues, incorrect citations, and increased risk of plagiarism, as well as inaccurate content, risk of excessive information leading to an infodemic on a particular topic, and cybersecurity issues [5].

A key question regarding the use of ChatGPT is if it can use evidence to identify premedical content. Evidence-based medicine (EBM) provides the highest level of evidence in medical treatment by integrating clinician experience, patient value, and best-available scientific information to guide decision-making on clinical management [7]. The principle of EBM means that the most appropriate treatment plan for patients should be devised based on the latest empirical research evidence. However, the scientific information identified by ChatGPT is not yet validated in terms of safety or accuracy according to Sallam [5], who further suggests that neither doctors nor patients should rely on it at this stage. In contrast, another study by Zhou et al [8] found that answers provided by ChatGPT were generally based on the latest verified scientific evidence, that is, the advice given followed high-quality treatment protocols and adhered to guidelines from experts.

In medicine, a clinical decision support system (CDSS) uses real-time evidence to support clinical decision-making. This is a fundamental tool in EBM, which uses SRs based on a systematic, scientific search of a particular subject. If ChatGPT becomes a CDSS, it is fundamental to determine whether it is capable of performing a systematic search based on real-time generation of evidence in the medical field. Therefore, this study will be the first to determine whether ChatGPT can search papers for an SR. In particular, this study aims to present a standard for medical research using generative AI search technology in the future by providing indicators for the reliability and consistency of generative AI searches from a user's perspective.

Methods

Ethical Considerations

As per 45 CFR §46.102(f), the activities performed herein were considered exempt from institutional review board approval due to the data being publicly available. Informed consent was not obtained, since this study used previously published deidentified information that was available to the general public. This study used publicly available data from PubMed, Embase, and Cochrane Library and did not include human participant research.

Setting the Benchmark

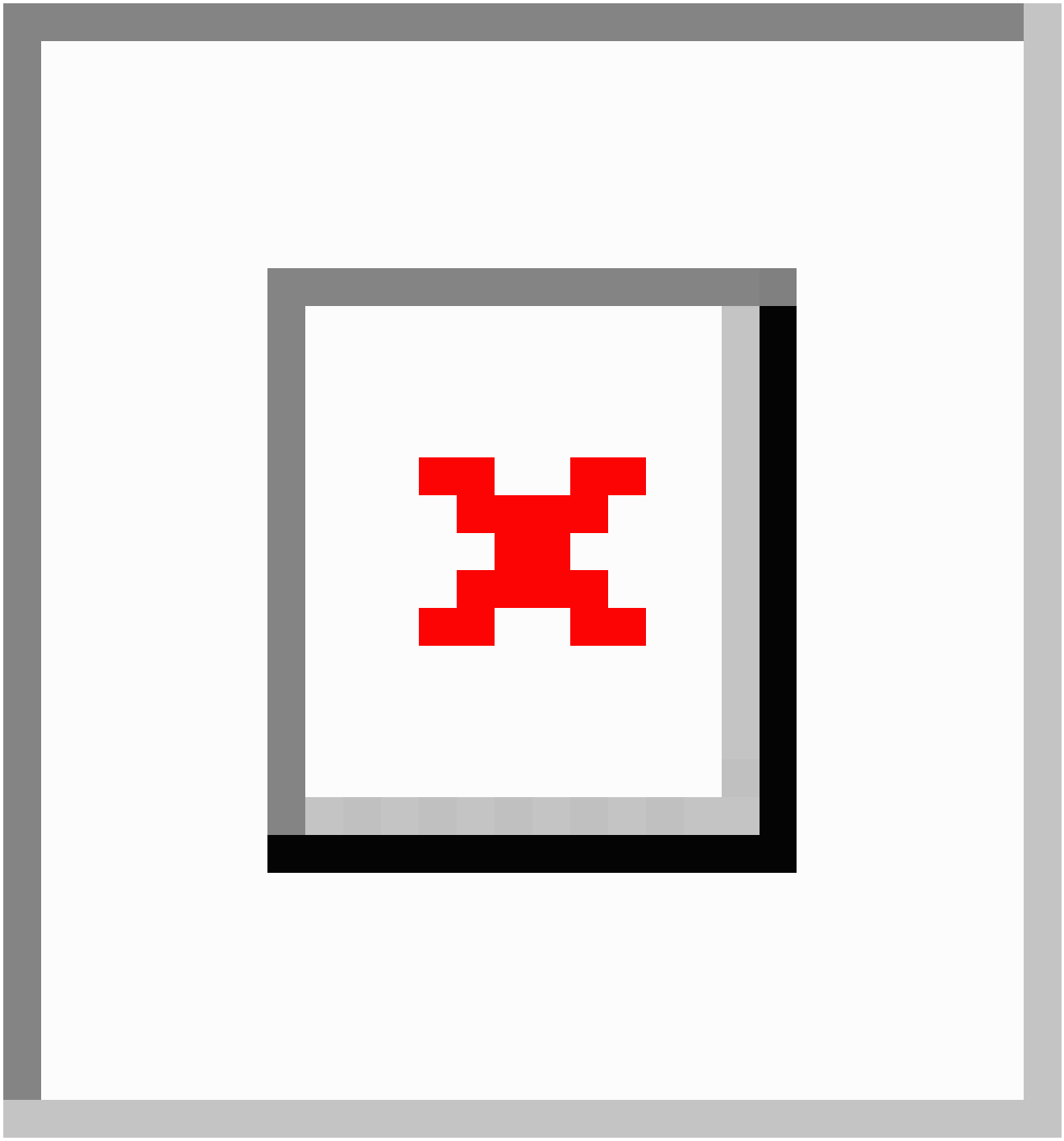
To determine whether ChatGPT, currently the most representative LLM, is capable of systematic searches, we set an SR that was performed by human experts as a benchmark and checked how many studies were finally included in the benchmark were presented by ChatGPT. We chose Lee et al [9] as the benchmark for the following reasons. First, Lee et al [9] performed an SR and meta-analysis about the medical treatment for Peyronie disease (PD) with human experts. PD typically presents with discomfort, curvature, or deformity of the penis in association with palpable plaques and erectile dysfunction [10]. Second, it was easy to compare the results of ChatGPT and the benchmark, because we had full information about the interim process and results of the study. Third, a sufficient amount of studies has been published about the medical treatment for PD, but there is still no consensus answer. So, we expected to assess the sole ability of ChatGPT as a systematic search tool with sufficient data while avoiding any possible pretrained bias. Lastly, with the topic of Lee et al [9], we could build questions that start broad and become more specific and add some conditions that could test ChatGPT's comprehension about scientific research. For example, questions could not only be built broadly by asking about "medical treatment for Peyronie's disease" but also specifically by asking about "oral therapy for Peyronie's disease" or "colchicine for Peyronie's disease." Because Lee et al [9] only contained randomized controlled trials (RCTs), we could add a condition

to the questions to restrict the study type to RCTs, which could be useful to assess the comprehension of ChatGPT.

Systematic Search Formula of Benchmark

Lee et al [9] used the following search query in PubMed and Cochrane Library: *("penile induration"[MeSH Terms] OR "Peyronie's disease"[Title/Abstract]) AND "male"[MeSH Terms] AND "randomized controlled trial"[Publication Type]*, and the following query in Embase: *('Peyronie disease'/exp OR 'Peyronie's diseas':ab,ti) AND 'male'/exp AND 'randomized controlled trial'/de*. After the systematic search, a total of 217 records were identified. Studies were excluded for the following reasons: not RCTs, not perfectly fit to the topic, not enough sample size or outcome, and not written in English. Finally, 24 RCTs were included in the SR, with only 1 RCT published in 2022 (Figure 1) [9]. The characteristics of all studies included in Lee et al [9] are summarized in Section S1 in Multimedia Appendix 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart for Lee et al [9]. RCT: randomized controlled trial.



Methodology of Systematic Search for ChatGPT

Based on the search formula used in Lee et al [9], a simple mandatory prompt in the form of a question was created, starting with comprehensive questions and gradually asking more specific questions (Textbox 1). For example, questions could be built as "Could you show RCTs of colchicine for Peyronie's disease in PubMed?" with the treatment and database changed

under the same format. In addition to mandatory questions, we added questions about treatment additionally provided by ChatGPT during the conversation. Considering the possibility that ChatGPT might respond differently depending on the interaction, we arranged questions into 2 logical flows, focusing on database and treatment, respectively (Figure 2 and Figure S1 in Multimedia Appendix 1). We asked about search results from 4 databases: PubMed [11], Google (Google Scholar) [12], Cochrane Library [13], and ClinicalTrials.gov [14]. PubMed is a leading biomedical database offering access to peer-reviewed articles. Google Scholar provides a wide-ranging index of scholarly literature, including medical studies. Cochrane Library specializes in high-quality evidence through SRs and clinical trials. ClinicalTrials.gov, managed by the National Library of Medicine, serves as a comprehensive repository for clinical study information globally. These databases collectively serve researchers by providing access to diverse and credible sources, facilitating literature reviews and evidence synthesis, and informing EBM in the medical field. They play crucial roles in advancing medical knowledge, supporting informed decision-making, and ultimately improving patient care outcomes [11-14]. These 4 databases were easy to access and contained most of the accessible studies. Each question was repeated at least twice. We extracted the answers and evaluated the quality of information based on the title, author, journal, and publication year (Sections S2-S5 Multimedia Appendix 1).

Textbox 1. Mandatory question prompts.

Basic format of questions

- "Could you show RCTs of (A) for Peyronie's disease in (B)?"

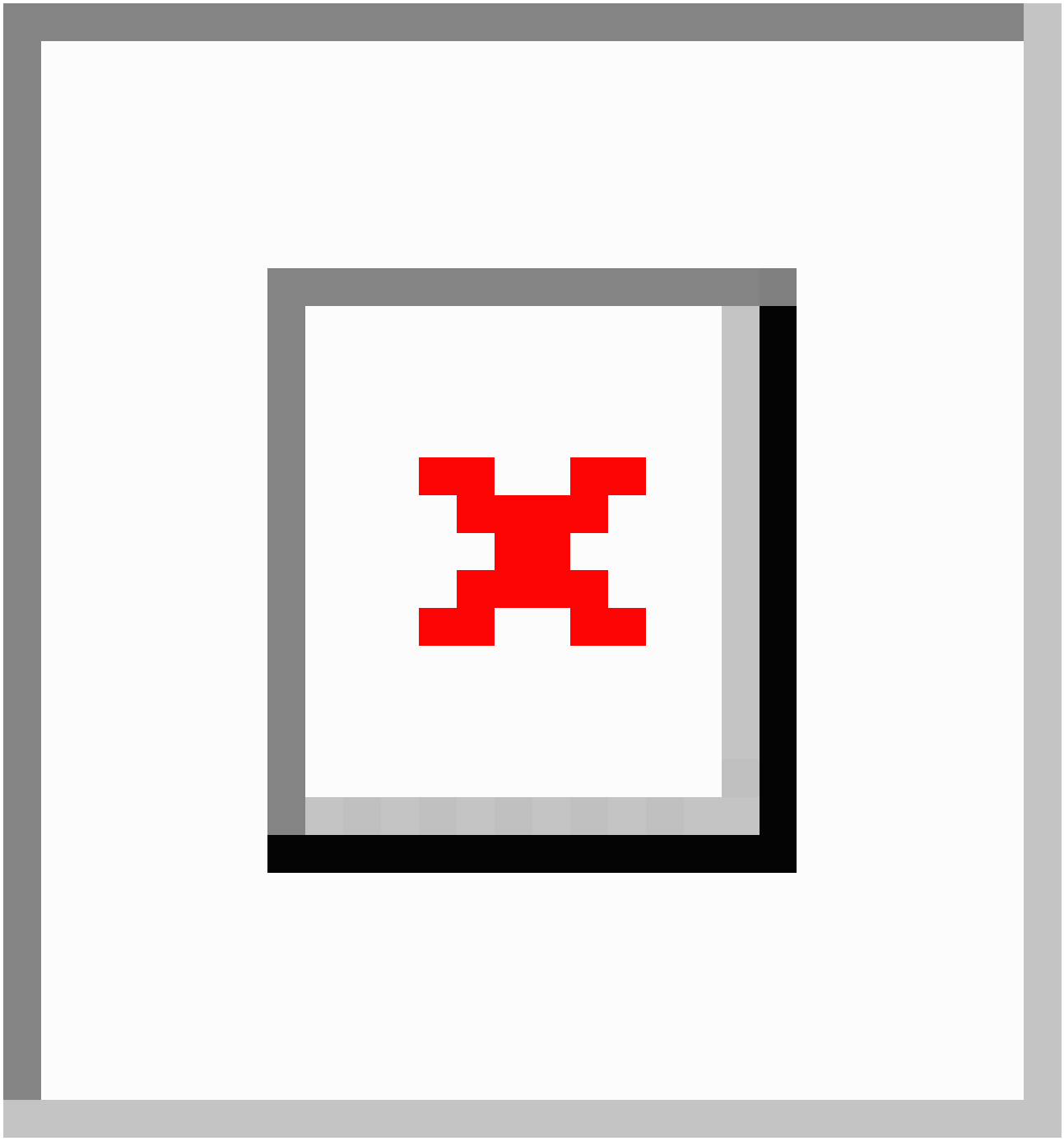
(A) Treatment category and specific treatment

- **Oral therapy**
  - Vitamin E, colchicine, L-carnitine, potassium aminobenzoate, tamoxifen, pentoxifylline, tadalafil, L-arginine, and sildenafil
- **Intralesional therapy**
  - Verapamil, interferon-a2B, collagenase *Clostridium histolyticum*, transdermal electromotive administration, hyaluronidase, triamcinolone, mitomycin C, super-oxide dismutase, and 5-fluorouracil
- **Mechanical therapy**
  - Extracorporeal shockwave therapy, iontophoresis, traction therapy, vacuum, penile massage, and exercise shockwave therapy
- **Topical therapy**
  - 5-Alpha-reductase inhibitors, superficial heat, diclofenac gel, collagenase *Clostridium histolyticum* gel, verapamil gel, potassium aminobenzoate gel, and propionyl-L-carnitine gel

(B) Database

- PubMed
- Google (Google Scholar)
- Cochrane Library
- ClinicalTrials.gov

Figure 2. Logical flow and results focusing on database for ChatGPT. RCT: randomized controlled trial.



We used the GPT-3.5 version of ChatGPT, which was pretrained with data before 2021, for the systematic search and evaluated how many RCTs that were included in Lee et al [9] were present in the search results from ChatGPT. To assess the reliability of ChatGPT's answers, we also evaluated whether the studies presented actually existed. ChatGPT's response style and the amount of information might vary from answer to answer. Thus, we evaluated the accuracy of the responses by prioritizing a match by (1) title; (2) author, journal, and publication year; and (3) other items.

To obtain higher-quality responses, it is important to structure the prompts using refined language that is well understood by the LLM [15-17]. In this study, we performed the following fine-tuning to clearly convey the most important content or information. We first defined roles and provided context and input data before asking complete questions to get responses, and we used specific and varied examples to help the model narrow its focus and produce more accurate results [18,19]. During the prompt engineering, the treatment category, specific treatment, and target databases were structured in order, and the order was changed in the detailed elements to induce consistent answers. Details of this are presented in Multimedia Appendix 1.

Quality Assessment of Answers

To evaluate the quality of individual studies derived from AI answers, we created a structured rating system based on bibliographic information related to the publications (Table 1).

We classified its answers into 4 grades if the title existed: A, B, C, and F. No grade was given for a fake title or no answer.

Table . Grade table based on bibliographic information.

Grade	Title actually exists	PICOS[a]	Essential information				Accessory information				Definition of grade
			Title	Author	Journal	Publication year	Issue number	Page number	DOI	PMID	
A	Yes	✓[b]	✓	✓	✓	✓	✓	✓	✓	✓	All bibliographic information matched
B	Yes	✓	✓	✓	✓	✓	Any X[c]	Any X	Any X	Any X	PICOS and essential information matched, but not accessory information
C	Yes	X[d]	✓	✓	✓	✓	N/A[e]	N/A	N/A	N/A	Essential information matched, but not PICOS
F	Yes	N/A	✓	Any X	Any X	Any X	N/A	N/A	N/A	N/A	Title matched, but not other essential information

[a]PICOS: population, intervention, comparison, outcome, and study design (research questions).

[b]Matched.

[c]Any mismatch in essential information or accessory information.

[d]Mismatch.

[e]N/A: not assessed.

A grade of "A" was given to an answer that was appropriate for the question and perfectly consistent with the actual study. For example, for the question "Could you show all RCTs of medical treatment for Peyronie's disease in Google?" ChatGPT answered "Gelbard MK, James K, Riach P, Dorey F. Collagenase versus placebo in the treatment of Peyronie's disease: a double-blind study. J Urol. 1993; 149(1):56-58." This answer was perfectly consistent with actual study because the authors, title, journal, publication year, issue number, and even page number matched, and it was appropriate because it was an RCT, was about medical treatment for PD, and could be found in Google.

A grade of "B" was given to an answer that was appropriate for the question and the title, author, journal, and publication year matched with the actual study. For example, for the question "Could you show RCTs of Intralesional therapy for Peyronie's disease in Cochrane Library?" ChatGPT answered "Gelbard MK, James K, Riach P, Dorey F. Collagenase versus placebo in the treatment of Peyronie's disease: a double-blind study. J Urol. 1993 Feb; 149(2):56-8. doi: 10.1016/s0022-5347(17)36148-1. PMID: 8426418." This answer was appropriate because it was an RCT, was about intralesional therapy for PD, and could be found in Cochrane Library. However, this answer was consistent with the title, authors, journal, and publication year of actual study but inconsistent with the accessory information for publication month, issue number, page number, DOI, and PMID. So, it was able to specify the presented study with essential information, but the accessory information was not consistent with the actual study.

A grade of "C" was given to an answer that was inappropriate for the question but the title, author, journal, and publication year matched with the actual study. For example, for the question "Could you show RCTs of Pentoxifylline for Peyronie's disease in PubMed?" ChatGPT answered "Gelbard MK, James K, Riach P, Dorey F. Collagenase versus placebo in the treatment of Peyronie's disease: a double-blind study. J Urol. 1993; 149(1):56-8. doi: 10.1016/s0022-5347(17)36091-7. PMID: 8417228." This answer was consistent with the title, authors, journal, and publication year of the actual study, but it was inappropriate because it was not about the use of pentoxifylline for PD.

A grade of "F" was given to an answer that was inconsistent with the author, journal, or publication year of the actual study that matched the title, making it impossible to be specified. Thus, it was considered as a fake study that was generated. For

example, for the question "Could you show RCTs of collagenase *Clostridium histolyticum* for Peyronie's disease in PubMed?" ChatGPT answered "Gelbard MK, James K, Riach P, Dorey FJ, & Collagenase Study Group. (2012). Collagenase versus placebo in the treatment of Peyronie's disease: a double-blind study. The Journal of urology, 187(3), 948-953." This answer was consistent with the title of the actual study but inconsistent with the authors, publication year, and so on.

Searching Strategy for Bing AI

To compare with ChatGPT, we performed the same process with Bing AI [20], also known as "New Bing," an AI chatbot developed by Microsoft and released in 2023. Since Bing AI functions based on the huge AI model "Prometheus" that includes OpenAI's GPT-4 with web searching capabilities, it is expected to give more accurate answers than the GPT-3.5 version of ChatGPT. We performed the conversation with the "Precise" tone. Because Bing AI limited the number of questions per session to 20, we did not arrange questions into 2 logical flows (Section S6 in Multimedia Appendix 1). We compared the number of studies included in the benchmark [9] and provided by Bing AI. We also evaluated the reliability of answers with the same method described above or using links of websites presented by Bing AI (Figure S2 and Section S7 in Multimedia Appendix 1).

Results

Systematic Search Results via ChatGPT

A total of 639 questions were entered into ChatGPT, and 1287 studies were obtained (Table 2). The systematic search via ChatGPT was performed from April 17 to May 6, 2023. At the beginning of the conversation, we gave ChatGPT the role of a researcher conducting a systematic search who intended to perform a meta-analysis for more appropriate answers. At first, we tried to build question format by using the word "find," such as "Could you find RCTs of medical treatment for Peyronie's disease?" However, ChatGPT did not present studies and only suggested how to find RCTs in a database, such as PubMed. Therefore, we changed the word "find" to "show," and ChatGPT presented lists of RCTs. For comprehensive questions, ChatGPT did not give an answer, saying that it did not have the capability to show a list of RCTs as an AI language model. However, when questions were gradually specified, it created answers (Sections S2 and S4 in Multimedia Appendix 1).

Table . Quality assessment of answers from ChatGPT and Bing AI[a].

Searcher, setting, and question level		Grade, n				Studies, n
		A	B	C	F	
ChatGPT						
	Database setting					
	Comprehensive question	1	0	3	5	56
	Category-specific question	1	1	8	18	124
	Treatment-specific question	4	7	67	87	545
	Total	6	8	78	110	725
	Treatment setting					
	Comprehensive question	0	0	0	1	27
	Category-specific question	0	0	4	8	61
	Treatment-specific question	1	10	85	92	474
	Total	1	10	89	101	562
	Total	7	18	167	211	1287
Bing AI						
	Comprehensive question	0	0	1	0	1
	Category-specific question	0	0	7	0	7
	Treatment-specific question	19	0	20	0	40
	Total	19	0	28	0	48
Human[b]		24	0	0	0	24

[a]AI: artificial intelligence.

[b]From Lee et al [9].

Of the 1287 studies provided by ChatGPT, only 7 (0.5%) studies were perfectly eligible and 18 (1.4%) studies could be considered suitable under the assumption that they were real studies if only the title, author, journal, and publication year matched (Table 2). Among these, only 1 study was perfectly consistent with studies finally included in Lee et al [9], and 4 studies were matched under the assumption (Sections S1, S3, and S5 in Multimedia Appendix 1).

Specifically, systematic search via ChatGPT was performed in 2 logical flow schemes, database setting and treatment setting (Figure 2 and Figure S1 in Multimedia Appendix 1). With the logical flow by database setting, among the 725 obtained studies, 6 (0.8%) and 8 (1.1%) studies were classified as grade A and grade B, respectively (Table 1). Of these, 1 grade A study and 1 grade B study were included in Lee et al [5]. With the logical flow by treatment setting, among the 562 obtained studies, 1 (0.2%) study was classified as grade A and 10 (1.8%) studies were classified as grade B. Of these, 3 grade B studies were included in the benchmark [9] (Table 2).

It was common for answers to be changed. There were many cases where answers contradicted themselves. In addition, there were cases where the answer was “no capability” or “no RCT found” at first, but when another question was asked and the previous question was asked again, an answer was given. ChatGPT showed a tendency to create articles by rotating some format and words. Titles presented were so plausible that it was almost impossible to identify fake articles until an actual search was conducted. The presented authors were also real people. Titles often contained highly specific numbers, devices, or brand names that were real. There were some cases where it was possible to infer which articles ChatGPT mimicked in the fake answers (Sections S3 and S5 in Multimedia Appendix 1). Considering these characteristics, when generating sentences, ChatGPT seemed to list words with a high probability of appearing among pretrained data rather than presenting accurate facts or understanding questions.

In conclusion, of the 1287 studies presented by ChatGPT, only 1 (0.08%) RCT matched the 24 RCTs of the benchmark [9].

Systematic Search Results via Bing AI

For Bing AI, a total of 223 questions were asked and 48 studies were presented. Among the 48 obtained studies, 19 (40%) studies were classified as grade A. There were no grade B

studies (Table 2). Because Bing AI always gave references with links to the websites, all studies presented by Bing AI existed. However, it also provided wrong answers about the study type, especially as it listed reviews as RCTs. Of the 28 studies with grade C, 27 (96%) were not RCTs and 1 (4%) was about a different treatment. Only 1 study had no grade because of a fake title; it presented a study registered in PubMed while pretending that it was the result of a search in ClinicalTrials.gov. However, the study was not in ClinicalTrials.gov (Section S7 in Multimedia Appendix 1).

Bing AI had more accurate answers than ChatGPT since it provides actual website references. However, it also showed a tendency to give more answers to more specific questions, similar to ChatGPT. For example, with a comprehensive question, Bing AI said "I am not able to access or search specific databases." However, with more specific questions, it found studies or answered "I couldn't find any RCTs' without mention about accessibility." In most cases, Bing AI either failed to find studies or listed too few studies to be used as a systematic searching tool.

In conclusion, of the 48 studies presented by Bing AI, 2 (4%) RCTs matched the 24 RCTs of the benchmark [9].

Discussion

Principal Findings

This paper's researchers sought to determine whether ChatGPT could conduct a real-time systematic search for EBM. For the first time, researchers compared the performance of ChatGPT with classic systematic searching as well as the Microsoft Bing AI search engine. Although Zhou et al [8] suggested that ChatGPT answered qualitative questions based on recent evidence, this study found that ChatGPT's results were not based on a systematic search (which is the basis for an SR), meaning that they could not be used for real-time CDSS in their current state.

With recent controversy regarding the risks and benefits of advanced AI technologies [21-24], ChatGPT has received mixed responses from the scientific community and academia. Although many scholars agree that ChatGPT can increase the efficiency and accuracy of the output in writing and conversational tasks [25], others suggest that the data sets used in ChatGPT's training might lead to possible bias, which not only limits its capabilities but also leads to the phenomenon of hallucination—apparently scientifically plausible yet factually inaccurate information [24]. Caution around the use of LLMs should also bear in mind security concerns, including the potential of cyberattacks that deliberately spread misinformation [25].

When applying the plug-in method in this study, especially when using PubMed Research [26], the process worked smoothly and there was not a single case of hallucination of fake research (by providing information along with a link), regardless of the designation of a specific database engine. Among the responses, 21 RCTs were included in the final SR, and out of a total of 24, all RCTs except 3 were provided. This is a very encouraging result. However, there is no plug-in that allows access to other databases yet, and if the conversation is long, the response speed is very slow. Furthermore, although it is a paid service, it only provides a total of 100 papers, so if more than 100 RCTs are searched, the user must manually search all papers. Ultimately, it is not intended for conducting an efficient and systematic search, as additional time and effort are required. If a more efficient plug-in is developed, this could play a promising part in systematic searches.

Although Sallam's [5] SR suggests that academic and scientific writing as well as health care practice, research, and education could benefit from the use of ChatGPT, this study found that ChatGPT could not search scientific articles properly, with a 0.08% (1/1287) of probability of the desired paper being presented. In the case of Bing AI using GPT-4, this study showed that Bing AI could search scientific articles with a much higher accuracy than ChatGPT. However, the probability was only 4% (2/48). It was still an insufficient probability for performing systematic research. Moreover, fake answers generated by ChatGPT, known as hallucinations, caused researchers to spend extra time and effort by checking the accuracy of the answers. A typical problem with generative AI is that it creates hallucinations. However, this is difficult to completely remove due to the principle of generative AI. Therefore, if it cannot be prevented from the pretraining of the model, efforts to increase reliability and consistency in the use of generative AI in medical care by checking the accuracy from the user's point of view are required, as shown in this study. Unlike ChatGPT, Bing AI did not generate fake studies. However, the total number of studies presented was too small. Very few studies have focused on the scientific searching accuracy of ChatGPT. Although this paper found many articles about the use of ChatGPT in the medical field, the majority concerned the role of ChatGPT as an author. Although the latter might accelerate writing efficiency, it also confirms the previously mentioned issues of transparency and plagiarism.

Wang et al [27] have recently investigated whether ChatGPT could be used to generate effective Boolean queries for an SR literature search. The authors suggest that ChatGPT should be considered a "valuable tool" for researchers conducting SRs, especially for time-constrained rapid reviews where trading off higher precision for lower recall is generally acceptable. They cite its ability to follow complex instructions and generate high-precision queries. Nonetheless, it should be noted that building a Boolean query is not a complex process. However, selecting the most appropriate articles for an SR is critical, which might be a more useful subject to examine in relation to the use of ChatGPT. Moreover, although Aydın and Karaarslan [28] have indicated that ChatGPT shows promise in generating a literature review, the iThenticate plagiarism tool found significant matches in paraphrased elements.

In scientific research, the most time-consuming and challenging task can be the process of filtering out unnecessary papers on the one hand and identifying those that are needed on the other hand. This difficult yet critical task can be daunting. It discourages many researchers from participating in scientific research. If AI could replace this process, it will be easier to collect and analyze data from the selected papers. Recently, commercial literature search services using generative AI models

have emerged. Representative examples include Covidence [29], Consensus [30], and Elicit [31]. The technical details of these commercial AI literature search services are unknown, but they are based on LLMs using GPT. Therefore, these search services are not only insufficient to verify hallucinations but also lack information in the search target databases. Even if there may be mistakes, the researcher should aim for completeness, and unverified methods should be avoided. Although this study did not use a commercial literature search service, it manually searched the target databases one by one. If the reliability and consistency of AI literature search services are verified, the use of these technologies will help medical research greatly

This study suggests that ChatGPT still has limitations in academic search, despite the recent assertion from Zhou et al [8] about its potential in searching for academic evidence. Moreover, although ChatGPT can search and identify guidance in open-access guidelines, its results are brief and fragmentary, often with just 1 or 2 sentences that lack relevant details about the guidelines.

Arguably, more concern should be placed on the potential use of ChatGPT in a CDSS than its role in education or writing draft papers. On the one hand, if AI such as ChatGPT is used within a patient-physician relationship, this is unlikely to affect liability since the advice is filtered through professionals' judgment and inaccurate advice generated by AI is no different from erroneous or harmful information disseminated by a professional. However, ChatGPT lacks sufficient accuracy and speed to be used in this manner. On the other hand, ChatGPT could also be used to give direct-to-consumer advice, which is largely unregulated since asking AI directly for medical advice or emotional support acts outside the established patient-physician relationship [32]. Since there is a risk of patient knowing inaccurate information, the medical establishment should seek to educate patients and guardians about the risk of inaccurate information in this regard.

Academic interest in ChatGPT to date has mainly focused on potential benefits including research efficiency and education, drawbacks related to ethical issues such as plagiarism and the risk of bias, as well as security issues including data privacy. However, in terms of providing medical information and acting as a CDSS, the use of ChatGPT is currently less certain because its academic search capability is potentially inaccurate, which is a fundamental issue that must be addressed.

The limitation of this study is that it did not address various research topics, because only 1 research topic was searched when collecting target literature. In addition, due to the time difference between the start of the study and the review and evaluation period, the latest technology could not be fully applied because it could become an outdated technology in a field of study where technology advances rapidly, such as generative AI. For example, there have already been significant technological advances since new AI models such as ChatGPT Turbo (4.0) were released between the time we started this study and the current revised time point.

This paper thus suggests that the use of AI as a tool for generating real-time evidence for a CDSS is a dream that has not yet become a reality. The starting point of evidence generation is a systematic search and ChatGPT is unsuccessful even for this initial purpose. Furthermore, its potential use in providing advice directly to patients in a direct-to-consumer form is concerning, since ChatGPT could provide inaccurate medical information that is not evidence based and can result in harm. For the proper use of generative AI in medical care in the future, it is suggested that a feedback model that evaluates accuracy according to experts' perspective, as done in this study, and then reflects it back into an LLM is necessary.

Conclusion

This is the first study to compare AI and conventional human SR methods as a real-time literature collection tool for EBM. The results suggest that the use of ChatGPT as a tool for real-time evidence generation is not yet accurate and feasible. Therefore, researchers should be cautious about using such AI. The limitations of this study using the GPT model are that the search for research topics was not diverse and that it did not prevent the hallucinations of generative AI. However, this study will serve as a standard for future studies by providing an index to verify the reliability and consistency of generative AI from a user's point of view. If the reliability and consistency of AI literature search services are verified, the use of these technologies will help medical research greatly.

Acknowledgments

This work was supported by the Soonchunhyang University Research Fund. This body had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Authors' Contributions

SRS had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. YNG, HSC, EJJ, JC, SL, and SRS contributed to the analysis and interpretation of data. YNG, HSC, SRS, and JHK contributed to the drafting of the manuscript. SRS and JHK contributed to critical revision of the manuscript for important intellectual content. YNG and SRS contributed to statistical analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional logical flow diagrams, characteristics of studies included in Lee et al [9], ChatGPT and Microsoft Bing transcripts, and grade classification for answers.
[DOCX File, 2209 KB - medinform_v12i1e51187_app1.docx]

References

1. Artificial intelligence (AI) in healthcare market (by component: software, hardware, services; by application: virtual assistants, diagnosis, robot assisted surgery, clinical trials, wearable, others; by technology: machine learning, natural language processing, context-aware computing, computer vision; by end user) - global industry analysis, size, share, growth, trends, regional outlook, and forecast 2022-2030. Precedence Research. 2023 Feb. URL: https://www.precedenceresearch.com/artificial-intelligence-in-healthcare-market [accessed 2024-03-31]
2. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 2021 Jul;8(2):e188-e194. [doi: 10.7861/fhj.2021-0095] [Medline: 34286183]
3. Zahlan A, Ranjan RP, Hayes D. Artificial intelligence innovation in healthcare: literature review, exploratory analysis, and future research. Technol Soc 2023 Aug;74:102321. [doi: 10.1016/j.techsoc.2023.102321]
4. Models. OpenAI. URL: https://platform.openai.com/docs/models/gpt-3-5 [accessed 2023-06-14]
5. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6):887. [doi: 10.3390/healthcare11060887] [Medline: 36981544]
6. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]
7. Evidence-Based Medicine Working Group. Evidence-based medicine. a new approach to teaching the practice of medicine. JAMA 1992 Nov 4;268(17):2420-2425. [doi: 10.1001/jama.1992.03490170092032] [Medline: 1404801]
8. Zhou Z, Wang X, Li X, Liao L. Is ChatGPT an evidence-based doctor? Eur Urol 2023 Sep;84(3):355-356. [doi: 10.1016/j.eururo.2023.03.037] [Medline: 37061445]
9. Lee HY, Pyun JH, Shim SR, Kim JH. Medical treatment for Peyronie's disease: systematic review and network Bayesian meta-analysis. World J Mens Health 2024 Jan;42(1):133. [doi: 10.5534/wjmh.230016]
10. Chung E, Ralph D, Kagioglu A, et al. Evidence-based management guidelines on Peyronie's disease. J Sex Med 2016 Jun;13(6):905-923. [doi: 10.1016/j.jsxm.2016.04.062] [Medline: 27215686]
11. PubMed. URL: https://pubmed.ncbi.nlm.nih.gov/about/ [accessed 2023-06-14]
12. Google Scholar. URL: https://scholar.google.com/ [accessed 2023-06-14]
13. Cochrane Library. URL: https://www.cochranelibrary.com/ [accessed 2023-06-14]
14. ClinicalTrials.gov. URL: https://classic.clinicaltrials.gov/ [accessed 2023-06-14]
15. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arXiv. Preprint posted online on Nov 28, 2023. [doi: 10.48550/arXiv.2311.16452]
16. Ziegler A, Berryman J. A developer's guide to prompt engineering and LLMs. GitHub Blog. 2023 Jul 17. URL: https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/ [accessed 2023-07-17]
17. Introducing ChatGPT. OpenAI. 2022 Nov 30. URL: https://openai.com/blog/chatgpt [accessed 2023-10-16]
18. Reid R. How to write an effective GPT-3 or GPT-4 prompt. Zapier. 2023 Aug 3. URL: https://zapier.com/blog/gpt-prompt/ [accessed 2023-10-14]
19. Prompt engineering for generative AI. Google. 2023 Aug 8. URL: https://developers.google.com/machine-learning/resources/prompt-eng?hl=en [accessed 2024-04-23]
20. Bing. URL: https://www.bing.com/ [accessed 2024-04-30]
21. de Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023 Apr 25;11:1166120. [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
22. Howard J. Artificial intelligence: implications for the future of work. Am J Ind Med 2019 Nov;62(11):917-926. [doi: 10.1002/ajim.23037] [Medline: 31436850]
23. Tai MCT. The impact of artificial intelligence on human society and bioethics. Tzu Chi Med J 2020 Aug 14;32(4):339-343. [doi: 10.4103/tcmj.tcmj_71_20] [Medline: 33163378]
24. Wogu IAP, Olu-Owolabi FE, Assibong PA, et al. Artificial intelligence, alienation and ontological problems of other minds: a critical investigation into the future of man and machines. Presented at: 2017 International Conference on Computing Networking and Informatics (ICCNI); Oct 29 to 31, 2017:; Lagos, Nigeria p. 1-10. [doi: 10.1109/ICCNI.2017.8123792]
25. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Frontiers in Computing and Intelligent Systems 2023 Jan 5;2(2):81-83. [doi: 10.54097/fcis.v2i2.4465]
26. PubMed Research. whatplugin.ai. URL: https://www.whatplugin.ai/plugins/pubmed-research [accessed 2024-04-30]
27. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? arXiv. Preprint posted online on Feb 9, 2023. [doi: 10.48550/arXiv.2302.03495]
28. Aydın Ö, Karaarslan E. OpenAI ChatGPT generated literature review: digital twin in healthcare. In: Aydın Ö, editor. Emerging Computer Technologies 2: İzmir Akademi Dernegi; 2022:22-31. [doi: 10.2139/ssrn.4308687]

29. Covidence. URL: https://www.covidence.org/ [accessed 2024-04-24]
30. Consensus. URL: https://consensus.app/ [accessed 2024-04-24]
31. Elicit. URL: https://elicit.com/ [accessed 2024-04-24]
32. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA 2023 Apr 25;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]

Abbreviations

AI: artificial intelligence
CDSS: clinical decision support system
EBM: evidence-based medicine
GPT: generative pre-trained transformer
LLM: large language model
PD: Peyronie disease
RCT: randomized controlled trial
SR: systematic review

Edited by A Castonguay; submitted 24.07.23; peer-reviewed by IG Jeong Jeong, J Noh, L Zhu, S Pandey, TG Rhee; revised version received 31.03.24; accepted 04.04.24; published 14.05.24.

Please cite as:
Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, Shim SR
The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation
JMIR Med Inform 2024;12:e51187
URL: https://medinform.jmir.org/2024/1/e51187
doi:10.2196/51187

Review

Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review

Yuan Peng[1], MSc; Franziska Bathelt[2], Dr Rer Nat; Richard Gebler[1], MSc; Robert Gött[3], Dipl.-Ing.; Andreas Heidenreich[4], Dipl.-Biol., Dr Rer Nat; Elisa Henke[1], MSc; Dennis Kadioglu[4,5], MSc; Stephan Lorenz[1], MSc; Abishaa Vengadeswaran[5], MSc; Martin Sedlmayr[1], Dr Rer Nat, Prof Dr

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

[2]Thiem-Research GmbH, Cottbus, Germany

[3]Core Unit Datenintegrationszentrum, University Medicine Greifswald, Greifswald, Germany

[4]Department for Information and Communication Technology (DICT), Data Integration Center (DIC), Goethe University Frankfurt, University Hospital, Frankfurt am Main, Germany

[5]Institute for Medical Informatics, Goethe University Frankfurt, University Hospital Frankfurt, Frankfurt am Main, Germany

Corresponding Author:
Yuan Peng, MSc
Institute for Medical Informatics and Biometry
Carl Gustav Carus Faculty of Medicine
Technische Universität Dresden
Fetscherstraße 74
Dresden, 01307
Germany
Phone: 49 3514583648
Fax: 49 3514585738
Email: yuan.peng@tu-dresden.de

Abstract

Background: Multisite clinical studies are increasingly using real-world data to gain real-world evidence. However, due to the heterogeneity of source data, it is difficult to analyze such data in a unified way across clinics. Therefore, the implementation of Extract-Transform-Load (ETL) or Extract-Load-Transform (ELT) processes for harmonizing local health data is necessary, in order to guarantee the data quality for research. However, the development of such processes is time-consuming and unsustainable. A promising way to ease this is the generalization of ETL/ELT processes.

Objective: In this work, we investigate existing possibilities for the development of generic ETL/ELT processes. Particularly, we focus on approaches with low development complexity by using descriptive metadata and structural metadata.

Methods: We conducted a literature review following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We used 4 publication databases (ie, PubMed, IEEE Explore, Web of Science, and Biomed Center) to search for relevant publications from 2012 to 2022. The PRISMA flow was then visualized using an R-based tool (Evidence Synthesis Hackathon). All relevant contents of the publications were extracted into a spreadsheet for further analysis and visualization.

Results: Regarding the PRISMA guidelines, we included 33 publications in this literature review. All included publications were categorized into 7 different focus groups (ie, medicine, data warehouse, big data, industry, geoinformatics, archaeology, and military). Based on the extracted data, ontology-based and rule-based approaches were the 2 most used approaches in different thematic categories. Different approaches and tools were chosen to achieve different purposes within the use cases.

Conclusions: Our literature review shows that using metadata-driven (MDD) approaches to develop an ETL/ELT process can serve different purposes in different thematic categories. The results show that it is promising to implement an ETL/ELT process by applying MDD approach to automate the data transformation from Fast Healthcare Interoperability Resources to Observational Medical Outcomes Partnership Common Data Model. However, the determining of an appropriate MDD approach and tool to implement such an ETL/ELT process remains a challenge. This is due to the lack of comprehensive insight into the characterizations of the MDD approaches presented in this study. Therefore, our next step is to evaluate the MDD approaches presented in this

study and to determine the most appropriate MDD approaches and the way to integrate them into the ETL/ELT process. This could verify the ability of using MDD approaches to generalize the ETL process for harmonizing medical data.

(JMIR Med Inform 2024;12:e52967) doi:10.2196/52967

KEYWORDS

ETL; ELT; Extract-Load-Transform; Extract-Transform-Load; interoperability; metadata-driven; medical domain; data harmonization

Introduction

Multisite clinical studies are increasingly using real-world data to gain real-world evidence, especially during the COVID-19 pandemic [1]. However, not all clinics use the same hospital information system, resulting in heterogeneity of data produced by different hospital information systems. These heterogeneous data are not semantically and syntactically interoperable. Therefore, it is difficult to analyze such data in a unified way across sites. For this, the heterogeneous data need to be harmonized and standardized, for example, by using a common data model (CDM) [2]. For example, the European Medical Agency [3] set up the DARWIN EU (Data Analysis and Real World Interrogation Network European Union) [4] to provide real-world evidence on use and adverse events of medicines across the European Union. DARWIN EU uses the Observational Medical Outcomes Partnership (OMOP) CDM [5] as the base model, which is provided by the Observational Health Data Sciences and Informatics [6] community. To participate in such networks, a transformation of local data is needed. A common approach is to develop an Extract-Transform-Load (ETL) or Extract-Load-Transform (ELT) process. Both are used to harmonize heterogeneous data into the target systems. The only difference between them is the order of processing data. ETL transforms the data before loading them into the target systems, while ELT loads the data into the target systems first, and then transforms the data. Due to the different data formats and source systems, multiple ETL/ELT processes have to be implemented [7-10]. This work is time-consuming and hard to maintain [11].

Using a standard data exchange format can reduce the complexity of transforming heterogeneous data into CDMs. An example is the Fast Healthcare Interoperability Resources (FHIR) [12] format. FHIR is a communication standard and is provided by the Health Level 7 (HL7) [13]. In Germany, the Medical Informatics Initiative (MII) [14] provides a Core Data Set (CDS) [15] in FHIR format for enabling the interoperability of data across all university hospitals. Another German association “the National Association of Statutory Health Insurance Physicians” (KBV, German: Kassenärztliche Bundesvereinigung) [16] also provides a KBV CDS in FHIR format, which provides a stable foundation for the development of the medical information objects [17] (eg, immunization records and maternity records). Although both MII CDS and KBV CDS are based on the German HL7 Basis Profiles [18], the FHIR profiles defined in the 2 CDSs are not identical [19]. This is due to the different requirements of MII and KBV. For example, codes indicating departments within a clinic (eg, 0100 for internal medicine department) are defined in different value-sets and therefore use different coding systems. This also complicates the implementation and maintenance of ETL/ELT processes.

Furthermore, most countries try to standardize their electronic health records (EHR) data for research and to improve the interoperability of the data. Consequently, country-specific FHIR profiles are developed, for example, German HL7 Basis Profiles [18] and the US CDS [20]. Due to different languages (ie, German vs English), different structure definitions (eg, extensions and cardinality) and different coding systems (eg, system URL for International Classification of Diseases, 10, Revision: German Modification [21] vs system URL for International Classification of Diseases, 10, Clinical Modification [22]) used in the FHIR profiles, different ETL processes need to be implemented [8,23]. Although these are just a few examples, it is conceivable that with the expansion of supported use cases, the time required for implementing an ETL/ELT process increases massively, while the maintainability decreases. Therefore, the implementation of a generic ETL/ELT process for harmonizing local health data can guarantee the semantic and syntactic interoperability of research data across sites and countries.

Using metadata for the implementation of ETL/ELT processes is a promising approach, as stated by David Loshin [24]: “in order to organize data for analytical purposes, it will need to be extracted from the original source (source metadata), transformed into a representation that is consistent with the warehouse (target metadata) in a way that does not lose information due to differences in format and precision (structure metadata) and is aligned in a meaningful way (semantic metadata).” A very broad definition of metadata is “data about other data” [25]. Depending on the specific context of use, metadata can be classified into 3 types [26]:

- **Descriptive metadata**: the metadata is used for discovery and identification purposes, for example metadata for source and target data.
- **Structural metadata**: the metadata is used for managing data in information systems, for example, column names and table names in a database.
- **Administrative metadata**: the metadata exists within a database that provides additional information, for example, the name of a person, who has changed the data in a database.

Metadata can be represented by metadata languages (eg, Resource Description Framework and Notation3) [27]. Such languages are also called ontology languages. For enabling the interoperability of data from different source and target systems, rule languages (eg, Rule Markup Language and Semantic Web

Rule Language) can be used to define the transformation rules between them [27]. Therefore, the use of metadata is expected to improve the development and maintenance for transforming FHIR resources to OMOP CDM.

As a side note, we understand any (descriptive and structural) metadata-based approach used for developing ETL/ELT processes as metadata-driven (MDD) approach. This work focuses on providing an overview of the types of MDD approaches and their use in different thematic categories. The overview aims to identify a suitable MDD approach to enhance the data transformation from FHIR to OMOP CDM. This will be achieved by answering the following questions:

- Q1: What are the themes of application for MDD approaches?
- Q2: What types of MDD approaches exist in the literature?
- Q3: What are the reasons for the usage of MDD approaches?
- Q4: What tool was used to implement the MDD approach?

Methods

To answer our 4 research questions, we conducted a literature review. To ensure the transparency of the review process, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [28]. We used 4 publication databases (ie, PubMed, IEEE Explore, Web of Science, and Biomed Center) to search for relevant publications from 2012 to 2022 written in German or English (Textbox 1). The first search was performed on August 11, 2022, and the second one was on March 15, 2023, which in turn completed the search through December 31, 2022. The collected publications were loaded into the Zotero Citation Management program (Corporation for Digital Scholarship) [29] and the duplicates were manually removed. To better categorize the publications to be excluded, we defined 8 exclusion criteria (Textbox 2).

This review was a 2-fold process consisting of Title-Abstract-Screening (TAS) and full-text screening (FTS). Both screening processes used the same exclusion criteria listed in Textbox 2. The unique publications were divided into 2 groups based on their publication dates and uploaded to a research collaboration platform, Rayyan (Qatar Computing Research Institute and Cochrane Bahrain) [30], as 2 separate projects. Each publication group was assigned with 4 reviewers. The corresponding author reviewed all publications. The TAS was performed under the blind-modus, so that each reviewer could label the publication independently. The blind-modus was turned off after all publications were tagged and the conflicts were discussed and resolved. After that, all included publications were randomly divided into 2 groups and reloaded into Rayyan as a new project for FTS. Similar to TAS, 4 reviewers were assigned to each publication group and the corresponding author reviewed all publications. The FTS was also conducted under the blind-modus and followed the same review process as the TAS.

We extracted the content of all included publications based on the categories listed in Textbox 3. The extraction of publication content was done by the corresponding author and validated by 4 coauthors. The extracted content was stored in a spreadsheet for further analysis and visualization.

The result of the literature review was visualized using an R-based tool, which was developed based on PRISMA 2020 [31].

Textbox 1. Search string and publication databases.

Search string

PubMed

- ((meta data) OR (meta-data) OR (metadata) OR (ontology) OR (rules)) AND ((extract transform load) OR (ETL) OR (extract load transform) OR (ELT))

IEEE Explore

- ((“All Metadata”:metadata) OR (“All Metadata”:meta-data) OR (“All Metadata”:meta data) OR (“All Metadata”:ontology) OR (“All Metadata”:rules)) AND ((“All Metadata”:ETL) OR (“All Metadata”:extract transform load) OR (“All Metadata”:ELT) OR (“All Metadata”:extract load transform))

Web of Science

- (ALL=(metadata) OR ALL=(meta-data) OR ALL=(“meta data”) OR ALL=(ontology) OR ALL=(rules)) AND (ALL=(ETL) OR ALL=(“extract transform load”) OR ALL=(ELT) OR ALL=(“extract load transform”))

Biomed Center (BMC)

- (“meta data” OR meta-data OR metadata OR ontology OR rules) AND (“extract transform load” OR ETL OR “extract load transform” OR ELT)

Textbox 2. Labels and descriptions of exclusion criteria.

Wrong_abbreviation

- Publication does not contain Extract-Transform-Load (ETL) as “Extract-Transform-Load.”
- Publication does not contain Extract-Load-Transform (ELT) as “Extract-Load-Transform.”

Wrong_definition

- Publication does not use metadata in the context of “metadata of data in source or target.”
- Publication does not use rules in the context of “rules for data transformation.”

Only_etl_elt

- Publication describes only ETL/ELT.

Only_metadata

- Publication describes only metadata.

Wrong_focus

- Publication mentioned metadata and ETL/ELT, but the focus is not about data harmonization

Wrong_type

- Publication is not a conference paper or a journal publication

Foreign_language

- Publication is written in other languages than English and German

Wrong_content

- Publication does not mention ETL/ELT or metadata

Textbox 3. Categories for data extraction.

Theme

- The main theme of the work.

Metadata-driven method

- The used metadata-driven method in the work.

Metadata-driven method tool

- Tool which was used to conduct the metadata-driven method.

Purpose

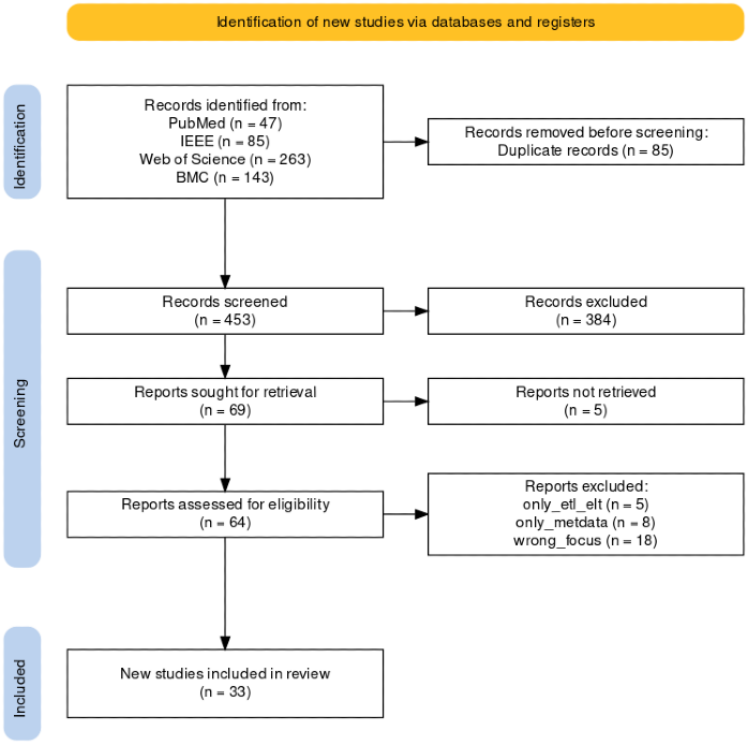
- The purpose of using the metadata-driven method.

Results

Literature Search

The literature search resulted in 538 publications. After removing 85 duplicates, 453 publications were screened during the TAS phase. By using the exclusion criteria defined in Textbox 2 and excluding the publications, which have no full-text, 64 publications were included for FTS. Finally, we included 33 publications in this work. The screening process and results are structured using the PRISMA flow diagram 2020 (Figure 1). A complete list of included publications is available in Multimedia Appendix 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. Generated using an R-based tool (reproduced from Haddaway et al [31], with permission from Neal R Haddaway).



Distribution of Publications

In order to gain an overview of the potential application focuses of MDDs (Q1) and thus an indication of where the approaches have proven beneficial, the focused theme of application was first evaluated. According to the extracted data, the focuses of all included publications are classified into 7 different categories, namely medicine (n=9) [10,32-39], data warehouse (n=13) [40-52], big data (n=4) [53-56], industry (n=4) [57-60], geoinformatics (n=1) [61], archaeology (n=1) [62], and military (n=1) [63]. This shows that data warehouse and medicine are the 2 categories that use the MDD approach the most.

MDD Approaches Used for Various Thematic Categories

Different types of MDD approaches were used across the thematic categories. To gain knowledge about the use of these types of MDD approaches in each category (Q2), the distribution of MDD approaches was investigated. Figure 2 shows the application of different types of MDD approaches in different thematic categories. The most frequently used type of MDD approach was ontology-based, where the ontology (using for example, resource description framework) of the source or target was applied in the ETL/ELT process. This approach was used in 6 categories, particularly in the categories of data warehouse [45-48,50,52] and medicine [10,32,35,37-39]. Another frequently used type of MDD approach was rule-based, which applied transformation rules generated based on the source and target to the ETL/ELT process. The rule-based approach was also widely used in the categories of data warehouse [40-43,49] and medicine [33,34,37,39]. All other MDD approaches besides the ontology-based and rule-based approaches were categorized as “other” (Table 1).

Figure 2. Metadata-driven approaches used in each thematic category.

Table 1. MDD[a] approaches that are categorized as "other."

MDD approach type and publication	Example
UML[b]-based	
Dhaouadi et al [46]	UML class diagram is used for modeling the transformation process
Graphic-based	
Dhaouadi et al [46]	BPMN[c] standard is used for modeling an ETL[d] process
Ad hoc formalisms-based	
Dhaouadi et al [46]	Entity Mapping Diagram is used for representing ETL tasks
MDA[e]-based	
Dhaouadi et al [46]	MDA is a multilayered framework with multiple submodules for separation of the specification of a functionality from its implementation
Message-based	
Novak et al [51]	"Normal message" contains information of mapping and transformation; "command message" configures the (execution) system
Template-based	
McCarthy et al [58]	A transformation template for each data source that manages the complex transformation process
Binding et al [62]	A template contains the mapping patterns which is then used for querying in database
Metadata-based[f]	
Ozyurt and Grethe [36]	Implementing a generic data transformation language to transform heterogeneous data from multiple sources to a common format
Tomingas et al [44]	Metadata of the source and target stored in a knowledge and metadata repository
Suleykin and Panfilov [60]	Metadata of the mapping path stored in a metadata management framework

[a]MDD: metadata-driven.

[b]UML: unified modeling language.

[c]BPMN: Business Process Model Notation.

[d]ETL: Extract-Transform-Load.

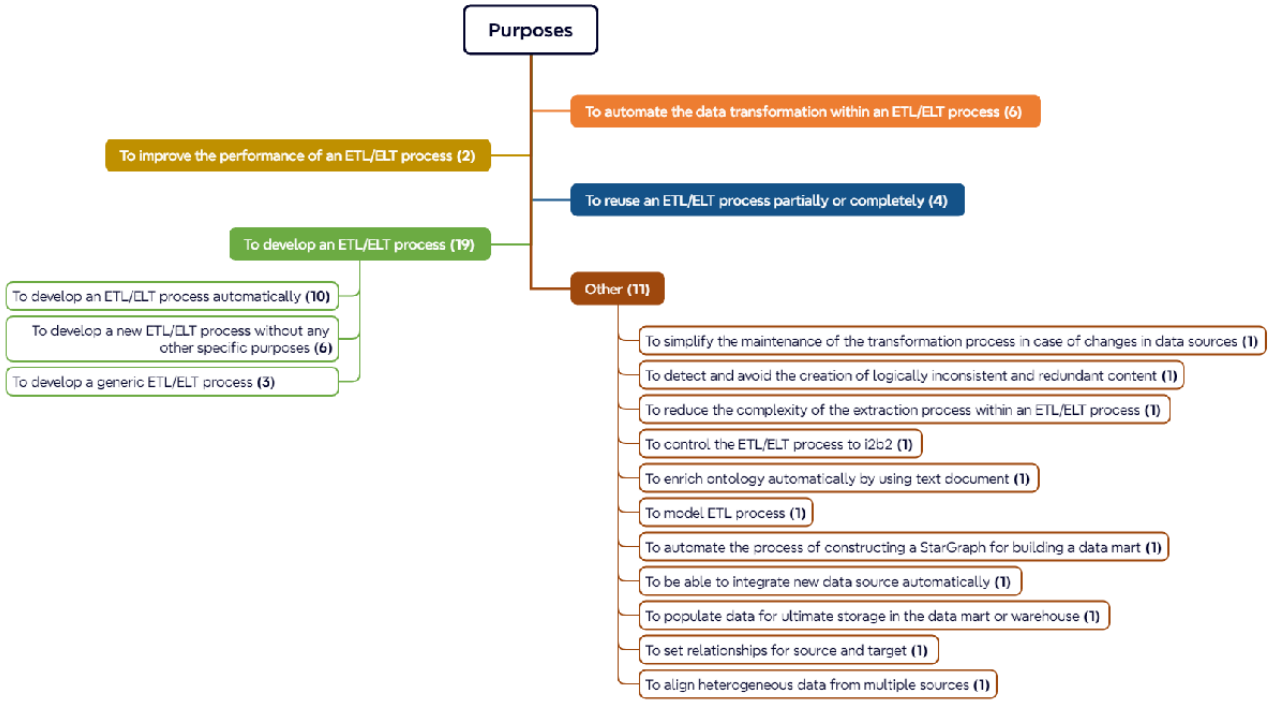
[e]MDA: Model Driven Architecture.

[f]Metadata-based approach: approach uses metadata without any specification.

Purposes of Using MDD Method for Data Harmonization

The purpose of using MDD approaches in each use case was then investigated to clarify the reasons why MDD approaches were used (Q3). Figure 3 shows different purposes of using MDD approaches in developing ETL/ELT processes based on the extracted data. The majority of publications describe the use of MDD approaches to develop an ETL/ELT process. This purpose can be divided into three detailed categories: (1) to automate the development of the ETL/ELT process [35,38,42,46,48-51,60], (2) to develop a generic ETL/ELT process [39,47,52], and (3) to develop a new ETL/ELT process without any further technical specifications [40,45,46,55,57,61]. Additionally, the transformation part of the ETL/ELT process could also be automated by applying an MDD approach [34,37,41,44,58,63]. For example, Chen and Zhao [41] described an MDD approach for the automatic generation of SQL scripts for data transformation. Moreover, using MDD approaches can also help to improve the performance of ETL/ELT processes [43,46] or to partially or fully reuse the ETL/ELT process [10,33,43,62]. Other goals (categorized as "Others" in Figure 3), such as simplifying the maintenance of the transformation process [37] and reducing the complexity of the extraction process [53], can also be realized by using MDD approaches in ETL/ELT processes.

Figure 3. Purposes of using MDD approaches in ETL/ELT process. ELT: Extract-Load-Transform; ETL: Extract-Transform-Load; i2b2: Informatics for Integrating Biology and the Bedside; MDD: metadata-driven.



Relationship Between Use Case and Used MDD Approach

As shown in the previous section, different MDD approaches were applied for different purposes. To further elucidate the reasons for choosing MDD approaches (Q3), the relationship between them was investigated. Table 2 lists the number of publications, which used a type of MDD approach to achieve a specific purpose. The ontology-based approach was used to achieve purposes (1) and (2), and (4)-(7). For example, Huang et al [63] created both local ontology (ontology based on the source data) and global ontology (ontology for the query processing) for the data transformation process, so that the data transformation from local ontology to global can be automated by applying ontology learning, ontology mapping, and ontology rules. Additionally, the ontology-based approach was also used to achieve other goals, such as controlling the ETL process to Informatics for Integrating Biology and the Bedside [32] and reducing the complexity of the extraction process [53]. Similar to the ontology-based approach, the rule-based approach was used to achieve the purposes of (1)-(3) and (5)-(7). Due to the reusability of the transformation rules, it was also possible to simplify the maintenance of the ETL/ELT process by applying rules in the process [37]. Other MDD approaches such as template-based [58,62], message-based [51], and metadata-based [41,44,48] were used to achieve the goals of (1)-(3) and (5)-(7). A metadata-based approach (eg, metadata management framework) can be used to develop the ETL tasks automatically [60]. The detailed information of Table 2 is available in the Multimedia Appendix 1.

Table 2. Relationships between purposes and MDD[a] approaches used.

Purposes		MDD approaches		
Number	Description	Ontology-based, n/N (%)	Rule-based, n/N (%)	Other, n/N (%)
(1)	To automate the data transformation within an ETL[b]/ELT[c] process	2/6 (33)	3/6 (50)	1/6 (17)
(2)	To reuse an ETL/ELT process (partially or completely)	1/4 (25)	2/4 (50)	1/4 (25)
(3)	To improve the performance of an ETL/ELT process	0/2 (0)	1/2 (50)	1/2 (50)
(4)	To develop a generic ETL/ELT process	3/3 (100)	0/3 (0)	0/3 (0)
(5)	To develop an ETL/ELT process automatically	5/9 (56)	2/9 (22)	2/9 (22)
(6)	To develop a new ETL/ELT process (without any other specific purposes)	4/6 (67)	1/6 (17)	1/6 (17)
(7)	Other	5/11 (45)	2/11 (18)	4/11 (36)

[a]MDD: metadata-driven.

[b]ETL: Extract-Transform-Load.

[c]ELT: Extract-Load-Transform.

Tools Used for Implementing MDD Approaches

Finally, we focused on the tools used to implemented MDD approaches (Q4). For achieving various purposes as shown in the previous section, different tools were used. As shown in Figure 4, each type of MDD approach can be implemented by using either an existing tool or a use case specific tool. Based on the included publications, the ontology-base approaches were mostly implemented using Protégé (Stanford Center for Biomedical Informatics Research) [64]. Protégé is an ontology editor, as well as OntoEdit (Institute AIFB, University of Karlsruhe and Ontoprise GmbH) [65]. The main reason for using an ontology editor is its ease of use and maintenance, as well as the various plug-ins. The use of case specific tools, such as ontology generator introduced by Kamil et al [45], generated ontologies based on the data definition language of the relational database. Both types of tools were used for creating and maintaining the ontology, which was then used to establish a generic mapping logic in the ETL/ELT process [32,50,52,54,55,61]. Another type of frequently used MDD approach is rule-based, which is used for phrasing and storing the transformation rules. The transformation rules can be stored in a mapping sheet [49], a CSV file [34], a YAML (YAML Ain't Markup Language) file [33] or a table within a database [43], which were implemented manually. Afterwards, the transformation rules could be used in the ETL/ELT process, for example, to enable the automatic transformation. Other types of MDD approaches can also be implemented by using existing tools (eg, knowledge and metadata repository [66]) or use case specific tools (eg, metadata repository [41] and metadata management framework [60]). For example, Ozyurt and Grethe [36] implemented a generic transformation language using the bioCADDIE Data Tag Suite (bioCADDIE Project) [67] (a metadata schema) to align heterogeneous data from multiple sources, which provided a basis for further analytic queries.

Figure 4. Tools used for developing the metadata-driven approach. MMF: metadata management framework; OWL: Web Ontology Language; YAML: YAML Ain't Markup Language.

Discussion

Principal Findings

Our literature review on the topic "metadata-driven ETL/ELT" includes all publications listed on PubMed, IEEE Explore, Web of Science, and Biomed Center on MDD ETL/ELT process from 2012 to 2022. In some context, the use of metadata is represented specifically using "ontology" or "rules." Therefore, we added "ontology" and "rules" into the search string to expand the search range.

With the review process presented, we were able to provide an overview of the thematic categories to which the MDD ETL/ELT processes were applied (Q1), the types of MDD approaches used in the ETL/ELT processes (Q2), the purposes of using MDD approaches (Q3), as well as the tools used to implement the MDD approaches (Q4).

Across all thematic categories, ontology-based and rule-based approaches are the most used approaches in the data warehouse and the medical thematic categories. In some cases, more than one MDD approach was used in the ETL/ELT process. For example, Del Carmen Legaz-García et al [39] used both ontology-based and rule-based approaches. Therefore, such publications were categorized as both MDD approach types.

Various tools can be used to implement MDD approaches. Unfortunately, we were not able to extract this information from all included publications. The reason for that is that some publications used proprietary or nontransferable approaches (eg, data-specific ontologies [39,62] and rules from Data Vault [DataVaultAlliance] [42]). Some other publications did not explicitly mention or describe the tools they used. Therefore, these publications were not included in the analysis of MDD tools used.

The results indicate that it is promising to implement a generic ETL/ELT process to transform different FHIR profiles to OMOP CDM automatically by applying MDD approaches. However, the results do not provide a trivial solution for this. For example, Huang et al [63] used an ontology-based approach to be able to automate the data transformation in an ETL/ELT process, while Ong et al [34] used a rule-based approach to achieve the same purpose. In some cases, more than one MDD approach were used as complements in order to accomplish the data transformation. For example, Pacaci et al [37] chose an ontology-based approach to automate the data transformation and a rule-based to simplify the maintenance of the transformation process in case of changes in data sources. By applying these 2 approaches in combination, the authors were able to transform EHR data from heterogeneous EHR systems into OMOP CDM. Therefore, determining an appropriate MDD approach and tool to implement a generic ETL/ELT process to transform FHIR to OMOP CDM automatically remains a challenge.

This work aimed to provide an overview of different types of MDD approaches and their tools. Consequently, this review lacks an analysis of detailing the specific traits of each MDD approach. This gap underscores the importance of providing a comprehensive insight into the characterizations of the MDD approaches presented in this study. This analysis will be conducted in the future to provide solid evidence for selecting the most suitable MDD approach and tool, or for considering using multiple MDD approaches in combination to implement the generic ETL/ELT process for transforming FHIR to OMOP CDM.

Conclusions

Our literature review shows that using MDD approaches to develop an ETL/ELT process can serve different purposes in different focus groups (ie, medicine, data warehouse, big data, industry, geoinformatics, archaeology, and military). The results show that it is promising to implement an ETL/ELT process by applying MDD approach for automating the data transformation from FHIR to OMOP CDM. However, the determination of an appropriate MDD approach and tool to implement such an ETL/ELT process remains a challenge. This is due to the lack of comprehensive insight into the characterizations of the MDD approaches presented in this study. Therefore, our next step is to evaluate the MDD approaches presented in this study and to determine the most appropriate MDD approaches and the way of integrating them into the MII CDS FHIR to OMOP CDM ETL process [8]. This could verify the ability of using MDD approaches to generalize the ETL process for harmonizing medical data [11].

Acknowledgments

This publication was partially funded by the German Federal Ministry of Education and Research (BMBF) Network of University Medicine 2.0: “NUM 2.0”, Grant No. 01KX2121, Project: NUM-Data integration center – NUM-DIZ. 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. YP did the search string definition and publications for the review-process preparation. YP, FB, Robert G, AH, EH, DK, SL, and AV: screened the title and abstract. YP, FB, Richard G, Robert G, AH, EH, DK, SL, and AV screened the full text. YP did the data extraction. FB, DK, Robert G, and SL performed the data extraction validation. YP wrote the original draft. YP, FB, Richard G, Robert G, AH, EH, DK, SL, AV, and MS reviewed and edited the writing. MS handled the resources. All authors have read and agreed to the current version of the paper and take responsibility for the scientific integrity of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excel tables for extracted data from included publications.
[XLSX File (Microsoft Excel File), 462 KB - medinform_v12i1e52967_app1.xlsx]

Multimedia Appendix 2
PRISMA-ScR checklist.
[DOCX File , 85 KB - medinform_v12i1e52967_app2.docx]

References

1. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022;22(1):287 [FREE Full text] [doi: 10.1186/s12874-022-01768-6] [Medline: 36335315]
2. 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]
3. European Medicines Agency. URL: https://www.ema.europa.eu/en [accessed 2022-08-18]
4. Data Analysis and Real World Interrogation Network (DARWIN EU). 2021. URL: https://www.darwin-eu.org/ [accessed 2023-12-16]
5. The Book of OHDSI: Observational Health Data Sciences and Informatics. San Bernardino, CA: OHDSI; 2019. URL: https://ohdsi.github.io/TheBookOfOhdsi [accessed 2024-01-19]
6. 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]

7. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the all of us research program: transforming i2b2 data into the OMOP common data model. PLoS One 2019;14(2):e0212463 [FREE Full text] [doi: 10.1371/journal.pone.0212463] [Medline: 30779778]
8. 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;169:104925 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104925] [Medline: 36395615]
9. Zoch M, Henke E, Reinecke I, Peng Y, Gebler R, Gruhl M, et al. Extract, transform and load German claim data to OMOP CDM—design and implications. Ger Medical Sci 2022:153 [FREE Full text] [doi: 10.3205/22gmds057]
10. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc 2016;23(5):909-915 [FREE Full text] [doi: 10.1093/jamia/ocv188] [Medline: 26911824]
11. Peng Y, Henke E, Sedlmayr M, Bathelt F. Towards ETL Processes to OMOP CDM Using Metadata and Modularization. Stud Health Technol Inform 2023;302:751-752 [FREE Full text] [doi: 10.3233/SHTI230256] [Medline: 37203486]
12. FHIR v4.0.1. HL7 International. URL: https://www.hl7.org/fhir/ [accessed 2022-04-05]
13. Kabachinski J. What is Health Level 7? Biomed Instrum Technol 2006;40(5):375-379 [FREE Full text] [doi: 10.2345/i0899-8205-40-5-375.1] [Medline: 17078369]
14. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
15. Ganslandt T, Boeker M, Löbe M, Prasser F, Schepers J, Semler SC, et al. Der Kerndatensatz der Medizininformatik-Initiative Ein Schritt zur Sekundärnutzung von Versorgungsdaten auf nationaler Ebene. Forum der Medizin-Dokumentation und Medizin-Informatik 2018;20(1):17-21.
16. The National Association of Statutory Health Insurance Physicians and the regional Associations of Statutory Health Insurance Physicians. Kassenärztliche Bundesvereinigung. 2020. URL: https://www.kbv.de/html/about_us.php [accessed 2023-08-01]
17. Medizinische Informationsobjekte (MIO). Kassenärztliche Bundesvereinigung. 2021. URL: https://www.kbv.de/html/mio.php [accessed 2023-08-01]
18. Leitfaden Basis DE (R4). HL7 FHIR Implementierungsleitfäden. URL: https://ig.fhir.de/basisprofile-de/stable/Home.html [accessed 2023-08-01]
19. Koch M, Richter J, Hauswaldt J, Krefting D. How to Make Outpatient Healthcare Data in Germany Available for Research in the Dynamic Course of Digital Transformation. Stud Health Technol Inform 2023;307:12-21 [FREE Full text] [doi: 10.3233/SHTI230688] [Medline: 37697833]
20. US Core implementation guide. HL7 International. URL: https://www.hl7.org/fhir/us/core/ [accessed 2022-12-16]
21. System URL for ICD-10-GM. Fast Healthcare Interoperability Resources. URL: http://fhir.de/CodeSystem/dimdi/icd-10-gm [accessed 2023-12-30]
22. System URL for ICD-10-CM. HL7 International. URL: http://hl7.org/fhir/sid/icd-10-cm [accessed 2023-12-30]
23. OMOPonFHIR Project. URL: https://omoponfhir.org/ [accessed 2022-04-05]
24. Loshin D. Chapter 9—metadata. In: Loshin D, editor. Business Intelligence: The Savvy Manager's Guide, 2nd Edition. Waltham, MA: Morgan Kaufmann; 2013:119-130.
25. Ulrich H, Kock-Schoppenhauer A, Deppenwiese N, Gött R, Kern J, Lablans M, et al. Understanding the nature of metadata: systematic review. J Med Internet Res 2022;24(1):e25440 [FREE Full text] [doi: 10.2196/25440] [Medline: 35014967]
26. ISO/IEC TR 19583-1:2019: information technology: concepts and usage of metadata—part 1: metadata concepts. International Organization for Standardization. 2019. URL: https://www.iso.org/standard/67365.html [accessed 2023-05-15]
27. Breitman KK, Casanova MA, Truszkowski W. Semantic Web: Concepts, Technologies and Applications. London: Springer; 2007.
28. 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;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
29. Zotero. 2022. URL: https://www.zotero.org/ [accessed 2022-02-10]
30. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
31. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. Campbell Syst Rev 2022;18(2):e1230 [FREE Full text] [doi: 10.1002/cl2.1230] [Medline: 36911350]
32. Post AR, Pai AK, Willard R, May BJ, West AC, Agravat S, et al. Metadata-driven clinical data loading into i2b2 for clinical and translational science institutes. AMIA Jt Summits Transl Sci Proc 2016;2016:184-193 [FREE Full text] [Medline: 27570667]
33. Quiroz JC, Chard T, Sa Z, Ritchie A, Jorm L, Gallego B. Extract, transform, load framework for the conversion of health databases to OMOP. PLoS One 2022;17(4):e0266911 [FREE Full text] [doi: 10.1371/journal.pone.0266911] [Medline: 35404974]

34. Ong TC, Kahn MG, Kwan BM, Yamashita T, Brandt E, Hosokawa P, et al. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. BMC Med Inform Decis Mak 2017;17(1):134 [FREE Full text] [doi: 10.1186/s12911-017-0532-3] [Medline: 28903729]
35. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H, Bürkle T, et al. Ontology-based data integration between clinical and research systems. PLoS One 2015;10(1):e0116656 [FREE Full text] [doi: 10.1371/journal.pone.0116656] [Medline: 25588043]
36. Ozyurt IB, Grethe JS. Foundry: a message-oriented, horizontally scalable ETL system for scientific data integration and enhancement. Database (Oxford) 2018;2018:bay130 [FREE Full text] [doi: 10.1093/database/bay130] [Medline: 30576493]
37. Pacaci A, Gonul S, Sinaci AA, Yuksel M, Erturkmen GBL. 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]
38. 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]
39. Del Carmen Legaz-García M, Miñarro-Giménez JA, Menárguez-Tortosa M, Fernández-Breis JT. Generation of open biomedical datasets through ontology-driven transformation and integration processes. J Biomed Semantics 2016;7:32 [FREE Full text] [doi: 10.1186/s13326-016-0075-z] [Medline: 27255189]
40. Gang H, Jin-Rong L, Xiu-Ying W. A kind of bidirectional mapping strategy of heterogeneous data model based on metadata-driven. 2012 Presented at: Proceedings of 2012 2nd International Conference on Computer Science and Network Technology; December 29-31, 2012; Changchun, China p. 1023-1027. [doi: 10.1109/iccsnt.2012.6526100]
41. Chen Z, Zhao T. A new tool for ETL process. 2012 Presented at: 2012 International Conference on Image Analysis and Signal Processing; November 09-11, 2012; Huangzhou, China p. 269-273. [doi: 10.1109/IASP.2012.6425038]
42. Puonti M, Raitalaakso T, Aho T, Mikkonen T. Automating transformations in data vault data warehouse loads. In: Thalheim B, Jaakkola H, Kiyoki Y, Yoshida N, editors. Information Modelling and Knowledge Bases XXVIII. Amsterdam: IOS Press; 2017:215-230.
43. Wang H, Zhang J, Guo J. Constructing data warehouses based on operational metadata-driven builder pattern. 2015 Presented at: 2015 International Conference on Logistics, Informatics and Service Sciences (LISS); July 27-29, 2015; Barcelona, Spain p. 1-4. [doi: 10.1109/liss.2015.7369630]
44. Tomingas K, Kliimask M, Tammet T. Data integration patterns for data warehouse automation. In: Vakali A, Trajcevski G, Kon-Popovska M, Ivanovic M, Bassiliades N, Palpanas T, et al, editors. New Trends in Database and Information Systems II. Berlin: Springer; 2015:41-55.
45. Kamil I, Inggriani MM, Asnar YDW. Data migration helper using domain information. 2014 Presented at: 2014 International Conference on Data and Software Engineering (ICODSE); November 26-27, 2014; Bandung, Indonesia p. 1-6. [doi: 10.1109/icodse.2014.7062492]
46. Dhaouadi A, Bousselmi K, Gammoudi MM, Monnet S, Hammoudi S. Data warehousing process modeling from classical approaches to new trends: main features and comparisons. Data 2022;7(8):113 [FREE Full text] [doi: 10.3390/data7080113]
47. Berkani N, Khouri S, Bellatreche L. Generic methodology for semantic data warehouse design: from schema definition to ETL. 2012 Presented at: 2012 Fourth International Conference on Intelligent Networking and Collaborative Systems; September 19-21, 2012; Bucharest, Romania p. 404-411. [doi: 10.1109/incos.2012.108]
48. Nath RPD, Romero O, Pedersen TB, Hose K. High-level ETL for semantic data warehouses. Semant Web 2022;13(1):85-132 [FREE Full text] [doi: 10.3233/sw-210429]
49. Yu QC. Metadata driven data mapper development. Appl Mech Mater 2013;411-414:403-407 [FREE Full text] [doi: 10.4028/www.scientific.net/amm.411-414.403]
50. Ta'a A, Abdullah MS. Ontology development for ETL process design. In: Ahmad MN, Abdullah MS, Colomb RM, editors. Ontology-based Applications for Enterprise Systems and Knowledge Management. Hershey, PA: Information Science Reference; 2013:261-275.
51. Novak M, Kermek D, Magdalenic I. Proposed architecture for ETL workflow generator. 2019 Presented at: Proceedings of the Central European Conference on Information and Intelligent Systems; October 2-4, 2019; Varaždin, Croatia p. 297-304.
52. Berkani N, Bellatreche L, Khouri S. Towards a conceptualization of ETL and physical storage of semantic data warehouses as a service. Cluster Comput 2013;16(4):915-931. [doi: 10.1007/s10586-013-0266-7]
53. Hilali I, Arfaoui N, Ejbali R. A new approach for integrating data into big data warehouse. 2022 Presented at: Proceedings Volume 12084, Fourteenth International Conference on Machine Vision (ICMV 2021); March 4, 2022; Rome, Italy p. 120841M. [doi: 10.1117/12.2623069]
54. Bansal SK, Kagemann S. Integrating big data: a semantic extract-transform-load framework. Computer 2015;48(3):42-50. [doi: 10.1109/mc.2015.76]
55. Bansal SK. Towards a semantic Extract-Transform-Load (ETL) framework for big data integration. 2014 Presented at: 2014 IEEE International Congress on Big Data; June 27-July 02, 2014; Anchorage, AK, USA p. 522-529. [doi: 10.1109/bigdata.congress.2014.82]

56. Boulahia C, Behja H, Louhdi MRC. Towards semantic ETL for integration of textual scientific documents in a big data environment: a theoretical approach. 2020 Presented at: 2020 6th IEEE Congress on Information Science and Technology (CiSt); June 05-12, 2021; Agadir-Essaouira, Morocco p. 133-138. [doi: 10.1109/cist49399.2021.9357280]
57. de Cesare C, Foy G, Lycett M. 4D-SETL a semantic data integration framework. 2016 Presented at: Proceedings of the 18th International Conference on Enterprise Information Systems—Volume 1: ICEIS; April 25-28, 2016; Rome, Italy p. 127-134. [doi: 10.5220/0005822501270134]
58. McCarthy S, McCarren A, Roantree M. A method for automated transformation and validation of online datasets. 2019 Presented at: 2019 IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC); October 28-31, 2019; Paris, France p. 183-189. [doi: 10.1109/edoc.2019.00030]
59. Scriney M, McCarthy S, McCarren A, Cappellari P, Roantree M. Automating data mart construction from semi-structured data sources. Comput J 2019;62(3):394-413. [doi: 10.1093/comjnl/bxy064]
60. Suleykin A, Panfilov P. Metadata-driven industrial-grade ETL system. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); December 10-13, 2020; Atlanta, GA, USA p. 2433-2442. [doi: 10.1109/bigdata50022.2020.9378367]
61. Janecka K, Cerba O, Jedlicka K, Jezek J. Towards interoperability of spatial planning Data: 5-Steps harmonization framework. 2013 Presented at: 13th SGEM GeoConference on Informatics, Geoinformatics and Remote Sensing; June 16-22, 2013; Albena, Bulgaria p. 1005-1016. [doi: 10.5593/SGEM2013/BB2.V1/S11.051]
62. Binding C, Charno M, Jeffrey S, May K, Tudhope D. Template based semantic integration: from legacy archaeological datasets to linked data. Int J Semantic Web Inf Syst 2015;11(1):1-29. [doi: 10.4018/ijswis.2015010101]
63. Huang DM, Du YL, Zhang MH, Zhang C. Application of ontology-based automatic ETL in marine data integration. 2012 Presented at: 2012 IEEE Symposium on Electrical & Electronics Engineering (EEESYM); June 24-27, 2012; Kuala Lumpur p. 11-13. [doi: 10.1109/eeesym.2012.6258574]
64. 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]
65. Sure Y, Angele J, Staab S. OntoEdit: multifaceted inferencing for ontology engineering. In: Aberer K, March S, Spaccapietra S, editors. Journal on Data Semantics I. LNCS 2800. Verlag Berlin: Springer; 2003:128-152.
66. MMX metadata framework. Mindworks Industries. URL: https://www.mindworks.industries/mmx_framework.html [accessed 2023-11-21]
67. Sansone SA, Gonzalez-Beltran A, Rocca-Serra P, Alter G, Grethe JS, Xu H, et al. DATS, the data tag suite to enable discoverability of datasets. Sci Data 2017;4:170059 [FREE Full text] [doi: 10.1038/sdata.2017.59] [Medline: 28585923]

Abbreviations

CDM: Common Data Model
CDS: Core Data Set
DARWIN EU: Data Analysis and Real World Interrogation Network European Union
EHR: electronic health record
ELT: Extract-Load-Transform
ETL: Extract-Transform-Load
FHIR: Fast Healthcare Interoperability Resources
FTS: full-text screening
HL7: Health Level 7
KBV: The National Association of Statutory Health Insurance Physicians (German: Kassenärztliche Bundesvereinigung)
MDD: metadata-driven
MII: Medical Informatics Initiative
OMOP: Observational Medical Outcomes Partnership
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TAS: Title-Abstract-Screening
YAML: YAML Ain’t Markup Language

Edited by C Lovis; submitted 20.09.23; peer-reviewed by M Löbe, W Xu; comments to author 24.10.23; revised version received 01.12.23; accepted 03.12.23; published 14.02.24.

Please cite as:
Peng Y, Bathelt F, Gebler R, Gött R, Heidenreich A, Henke E, Kadioglu D, Lorenz S, Vengadeswaran A, Sedlmayr M
Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review
JMIR Med Inform 2024;12:e52967
URL: https://medinform.jmir.org/2024/1/e52967
doi:10.2196/52967
PMID:38354027

Review

The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review

Carl Preiksaitis[1], MD; Nicholas Ashenburg[1], MD; Gabrielle Bunney[1], MBA, MD; Andrew Chu[1], MD; Rana Kabeer[1], MPH, MD; Fran Riley[1], MSE, MD; Ryan Ribeira[1], MPH, MD; Christian Rose[1], MD

Department of Emergency Medicine, Stanford University School of Medicine, Palo Alto, CA, United States

Corresponding Author:
Carl Preiksaitis, MD
Department of Emergency Medicine
Stanford University School of Medicine
900 Welch Road
Suite 350
Palo Alto, CA, 94304
United States
Phone: 1 650 723 6576
Email: cpreiksaitis@stanford.edu

Abstract

Background: Artificial intelligence (AI), more specifically large language models (LLMs), holds significant potential in revolutionizing emergency care delivery by optimizing clinical workflows and enhancing the quality of decision-making. Although enthusiasm for integrating LLMs into emergency medicine (EM) is growing, the existing literature is characterized by a disparate collection of individual studies, conceptual analyses, and preliminary implementations. Given these complexities and gaps in understanding, a cohesive framework is needed to comprehend the existing body of knowledge on the application of LLMs in EM.

Objective: Given the absence of a comprehensive framework for exploring the roles of LLMs in EM, this scoping review aims to systematically map the existing literature on LLMs' potential applications within EM and identify directions for future research. Addressing this gap will allow for informed advancements in the field.

Methods: Using PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) criteria, we searched Ovid MEDLINE, Embase, Web of Science, and Google Scholar for papers published between January 2018 and August 2023 that discussed LLMs' use in EM. We excluded other forms of AI. A total of 1994 unique titles and abstracts were screened, and each full-text paper was independently reviewed by 2 authors. Data were abstracted independently, and 5 authors performed a collaborative quantitative and qualitative synthesis of the data.

Results: A total of 43 papers were included. Studies were predominantly from 2022 to 2023 and conducted in the United States and China. We uncovered four major themes: (1) clinical decision-making and support was highlighted as a pivotal area, with LLMs playing a substantial role in enhancing patient care, notably through their application in real-time triage, allowing early recognition of patient urgency; (2) efficiency, workflow, and information management demonstrated the capacity of LLMs to significantly boost operational efficiency, particularly through the automation of patient record synthesis, which could reduce administrative burden and enhance patient-centric care; (3) risks, ethics, and transparency were identified as areas of concern, especially regarding the reliability of LLMs' outputs, and specific studies highlighted the challenges of ensuring unbiased decision-making amidst potentially flawed training data sets, stressing the importance of thorough validation and ethical oversight; and (4) education and communication possibilities included LLMs' capacity to enrich medical training, such as through using simulated patient interactions that enhance communication skills.

Conclusions: LLMs have the potential to fundamentally transform EM, enhancing clinical decision-making, optimizing workflows, and improving patient outcomes. This review sets the stage for future advancements by identifying key research areas: prospective validation of LLM applications, establishing standards for responsible use, understanding provider and patient perceptions, and improving physicians' AI literacy. Effective integration of LLMs into EM will require collaborative efforts and thorough evaluation to ensure these technologies can be safely and effectively applied.

(JMIR Med Inform 2024;12:e53787) doi:10.2196/53787

KEYWORDS

large language model; LLM; emergency medicine; clinical decision support; workflow efficiency; medical education; artificial intelligence; AI; natural language processing; NLP; AI literacy; ChatGPT; Bard; Pathways Language Model; Med-PaLM; Bidirectional Encoder Representations from Transformers; BERT; generative pretrained transformer; GPT; United States; US; China; scoping review; Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA; decision support; workflow efficiency; risk; ethics; education; communication; medical training; physician; health literacy; emergency care

Introduction

Background

Emergency medicine (EM) is at an inflection point. With increasing patient volumes, decreasing staff availability, and rapidly evolving clinical guidelines, emergency providers are overburdened and burnout is significant [1]. While the role of artificial intelligence (AI) in enhancing emergency care is increasingly recognized, the emergence of large language models (LLMs) offers a novel perspective. Previous reviews have systematically categorized AI applications in EM, focusing on diagnostic-specific and triage-specific branches, emphasizing diagnostic prediction and decision support [2-5]. This review aims to build upon these foundations by exploring the unique potential of LLMs in EM, particularly in areas requiring complex data processing and decision-making under time constraints.

An LLM is a deep learning–based artificial neural network, distinguished from traditional machine learning models by its training on vast amounts of textual data. This enables LLMs to recognize, translate, predict, or generate text or other content [6]. Characterized by transformer architecture and the ability to encode contextual information using several parameters, LLMs allow for nuanced understanding and application across a diverse range of topics. Unlike traditional AI models, which often rely on structured data and predefined algorithms, LLMs are adept at interpreting unstructured text data. This feature makes them particularly useful in tasks such as real-time data interpretation, augmenting clinical decision-making, and enhancing patient engagement in clinical settings. For instance, LLMs can efficiently sift through electronic health records (EHRs) to identify critical patient histories and assist clinicians in interpreting multimodal diagnostic data. In addition, they can serve as advanced decision support tools in differential diagnosis, enhancing the quality of care while reducing the cognitive load and decision fatigue for emergency providers. Furthermore, the content generation ability of LLMs, ranging from technical computer code to essays and poetry, demonstrates their versatility and exceeds the functional scope of traditional machine learning models in terms of content creation and natural language processing.

Importance

While interest in applying LLMs to EM is gaining momentum, the existing body of literature remains a patchwork of isolated studies, theoretical discussions, and small-scale implementations. Moreover, existing research often focuses on specific use cases, such as diagnostic assistance or triage prioritization, rather than providing a holistic view of how LLMs can be integrated into the EM workflow. Conclusions based on other forms of machine learning are not readily translatable to LLMs. This fragmented landscape makes it challenging for emergency clinicians, who are already burdened by the complexities and pace of their practice, to discern actionable insights or formulate a coherent strategy for adopting these technologies. Despite the promise shown by several models, such as ChatGPT-4 (OpenAI) or Med-PaLM 2 (Google AI), the absence of standardized metrics for evaluating their clinical efficacy, ethical use, and long-term sustainability leaves researchers and clinicians navigating an uncharted territory. Consequently, the potential for LLMs to enhance emergency medical care remains largely untapped and poorly understood.

Goals of This Review

In light of these complexities and informational disparities, our study undertakes a crucial step to consolidate, assess, and contextualize the fragmented knowledge base surrounding LLMs in EM. Through a scoping review, we aim to establish a foundational understanding of the field's current standing, from technological capabilities to clinical applications and ethical considerations. This synthesis serves a dual purpose: first, to equip emergency providers with a navigable map of existing research and, second, to identify critical gaps and avenues for future inquiry. As EM increasingly embraces technological solutions for its unique challenges, our goal is to provide clarity to the responsible and effective incorporation of LLMs into clinical practice.

Methods

Overview

We adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist [7] and used the scoping review methodology proposed by Arksey and O'Malley [8] and furthered by Levac et al [9]. This included the following steps: (1) identifying the research question; (2) identifying relevant studies; (3) selecting studies; (4) charting the data; (5) collating, summarizing, and reporting the results; and (6) consultation. Our full review protocol is published elsewhere [10].

Identifying the Research Question

The overall purpose of this review was to map the current literature describing the potential uses of LLMs in EM and to identify directions for future research. To achieve this goal, we aimed to answer the primary research question: "What are the current and potential uses of LLMs in EM described in the literature?" We chose to explicitly focus on LLMs as this subset of AI is rapidly developing and generating significant interest for potential applications.

Identifying Relevant Studies

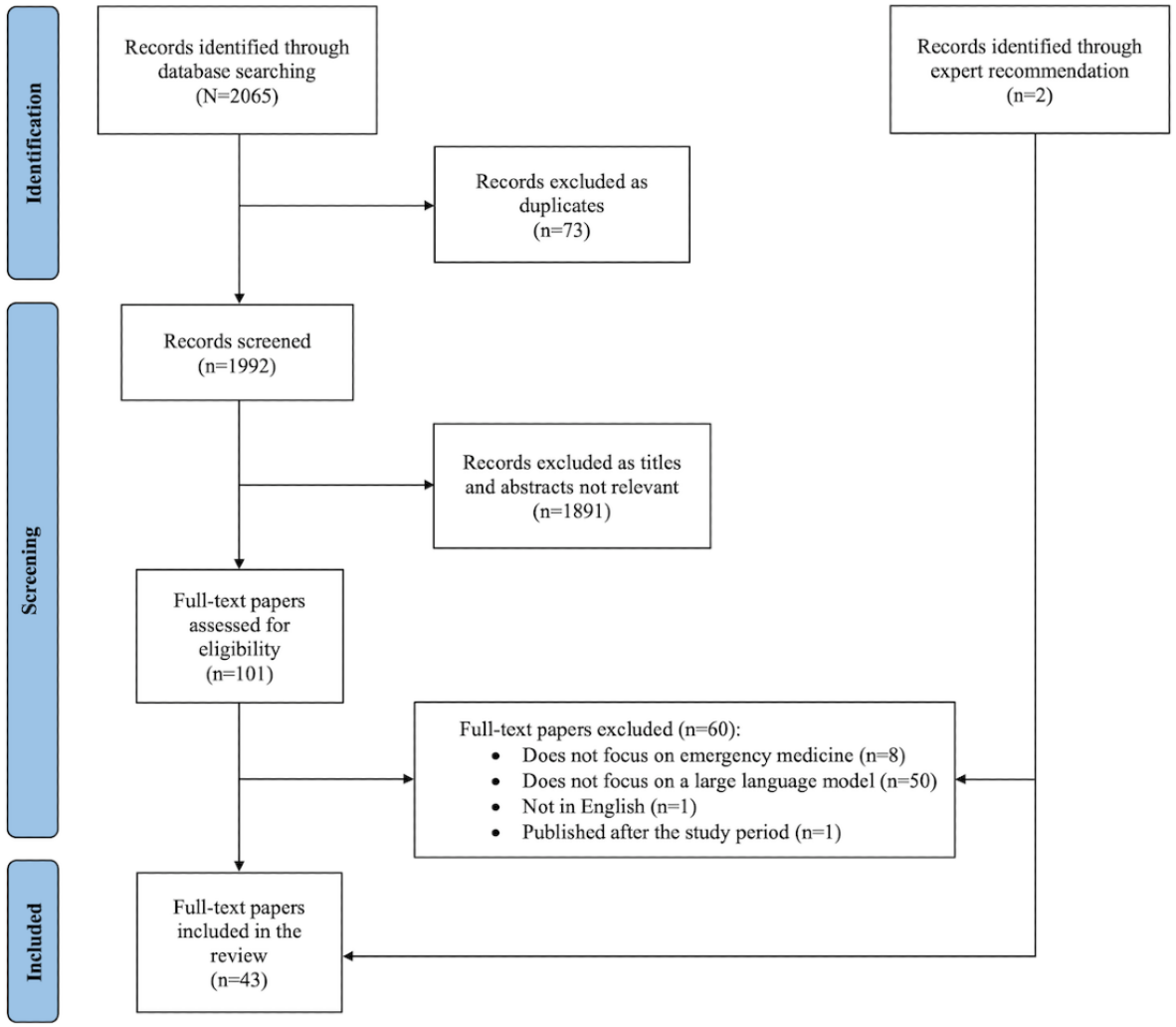
In August 2023, we searched Ovid MEDLINE, Embase, Web of Science, and Google Scholar for potential citations of interest. We limited our search to papers published after January 2018 as the Bidirectional Encoder Representations from Transformers (BERT; Google) model was introduced that year and considered by many to be the first in the contemporary class of LLMs [11]. Our search strategy (Multimedia Appendix 1), created in consultation with a medical librarian, combined keywords and MeSH (Medical Subject Headings) terms related to LLMs and EM. We reviewed the bibliographies of identified studies for potential missed papers.

Study Selection

Citations were managed using Covidence web-based software (Veritas Health Innovation). Manuscripts were included if they discussed the use of an LLM in EM, including applications in the emergency department (ED) and prehospital and periadmission settings. Furthermore, we included use cases related to public health, disease monitoring, or disaster preparedness as these are relevant to EDs. We excluded studies that used other forms of machine learning or natural language processing that were not LLMs and studies that did not clearly relate to EM. We also excluded cases where the only use of an LLM was in generating the manuscript without any additional commentary.

Two investigators (CP and CR) independently screened 100 abstracts, and the interrater reliability showed substantial agreement (κ=0.75). The remaining abstracts were screened by 1 author (CP), who consulted with a second author as needed for clarification regarding inclusion and exclusion criteria. All papers meeting the initial criteria were independently reviewed in full by 2 authors (CP and CR). Studies determined to meet the eligibility criteria by both reviewers were included in the analysis. Discrepancies were resolved by consensus and with the addition of a third reviewer (NA) if needed. Our initial search strategy identified 2065 papers, of which 73 (3.54%) were duplicates, resulting in 1992 (96.46%) papers for screening (Figure 1). Of the 1992 papers, 1891 (94.93%) were excluded based on the title or abstract. In total, 5.07% (101/1992) of the papers were reviewed in full, and 2.11% (42/1992) of the papers were found to meet the study inclusion criteria. During manuscript review, 2 additional papers were brought to our attention by experts, and 1 of these met the inclusion criteria, bringing the total number of included papers to 43.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of search and screening for large language models in emergency medicine.



Charting the Data

Data abstraction was independently conducted using a structured form to capture paper details, including the author, year of publication, study type, specific study population, study or paper location, purpose, and main findings. Data to address our primary research question was iteratively abstracted from the papers as our themes emerged, as explained in the subsequent sections.

Collating, Summarizing, and Reporting the Results

We synthesized and collated the data, performing both a quantitative and qualitative analysis. A descriptive summary of the included studies was created. Then, we used the methodology proposed by Braun and Clarke [12] to conduct a thematic analysis to address our primary research question. Five authors (CP, CR, AC, NA, and RR) independently familiarized themselves with and generated codes for a purposively diverse selection of 10 papers, focusing on content that suggested possible uses for LLMs in EM. The group met to discuss preliminary findings and refine the group's approach. Individuals then independently aggregated codes into themes. These themes were reviewed and refined as a group. Then, 2 authors (CP and CR) reviewed the remaining manuscripts for any additional themes and data that supported or contradicted our existing themes. These data were used to refine themes through group discussion. Our analysis included a discussion and emphasis on the implications and future research directions for the field, based on the guidance from Levac et al [9].

Consultation

To ensure our review accurately characterized the available knowledge and that our interpretations of it were correct, we consulted with external emergency physicians with topic expertise in AI. We incorporated feedback as appropriate. For example, we more completely defined LLMs for clarity and included a table describing common models (Table 1). Our findings and recommendations were endorsed by our consultants.

Table 1. Large language models reported in the identified literature.

Model	Interface	Model size (parameters)	Developer	Year of release
GPT-3.5 Turbo	ChatGPT	175 billion [13]	OpenAI	2022
GPT-4	ChatGPT	Approximately 1.8 trillion (estimated) [14]	OpenAI	2023
Pathways Language Model	Bard	540 billion [15]	Google AI[a]	2023
Embeddings from Language Model	Full model available	93.6 billion [16]	Allen Institute for AI	2018
Bidirectional Encoder Representations from Transformers	Full model available	110 million and 340 million [17]	Google	2018

[a]AI: artificial intelligence.

Results

Overview

Most identified studies (29/43, 67%) were published in 2023. Of the 43 studies, 14 (33%) were conducted in the United States, followed by 6 (14%) in China, 4 (9%) in Australia, 3 (7%) each in Taiwan and France, and 2 (5%) each in Singapore and Korea. Several other individual studies (5/43, 12%) were from various countries (Table 2).

In terms of study type, 40% (17/43) of the papers were methodology studies; 40% (17/43) were case studies; 16% (7/43) were commentaries; and 2% (1/43) each of a case report, qualitative investigation, and retrospective cross-sectional study. In total, 58% (25/43) of these studies addressed the ED setting specifically, followed by 14% (6/43) addressing the prehospital setting and 14% (6/43) addressing other non-ED hospital settings. In total, 7% (3/43) of the studies focused on using LLMs for the public, 5% (2/43) focused on using them for social media analysis, and 2% (1/43) focused on using them for research applications. LLMs used in the reviewed papers (Table 1) included versions of GPT (OpenAI; eg, ChatGPT, GPT-4, and GPT-2), Pathways Language Model (Bard; Google AI), Embeddings from Language Model, XLNet, and BERT (Google; eg, BioBERT, ClinicalBERT, and decoding-enhanced BERT with disentangled information).

We identified four major themes in our analysis: (1) clinical decision-making and support; (2) efficiency, workflow, and information management; (3) risks, ethics, and transparency; and (4) education and communication. Major themes, subthemes, and representative quotations are presented in Table 3.

Table 2. Summary of included studies and identified themes (N=43).

Study	Country	Study type	Purpose	Setting and context	Large language models used	Sample size	Themes
Xu et al [18], 2020	France	Methodology	Classification of visits into trauma and nontrauma based on ED[a] notes	ED	GPT-2 (OpenAI)	16,1930 notes	CDMS[b] and EWIM[c]
Wang et al [19], 2020	China	Retrospective cross-sectional study	Sentiment analysis of social media posts related to COVID-19	Social media	BERT[d] (Google)	99,9978 posts	EWIM
Chen et al [20], 2020	Taiwan	Methodology	Diagnosis identification from discharge summaries	Inpatient	BERT and BioBERT	25,8850 discharge diagnoses	EWIM
Chang et al [21], 2020	United States	Methodology	Categorize free-text ED chief complaints	ED	BERT and Embeddings from Language Model	2.1 million adult and pediatric ED visits	CDMS and EWIM
Wang et al [22], 2021	Singapore	Methodology	Summarize EMS[e] reports for clinical audits	EMS and prehospital	BERT	58,898 ambulance incidents	EWIM
Gil-Jardiné et al [23], 2021	France	Methodology	Classify content of EMS calls during the COVID-19 pandemic	EMS and prehospital	GPT-2	888,469 calls (training), 39,907 calls (validation), and 254,633 calls (application)	EWIM
Shung et al [24], 2021	United States	Methodology	Identify patients with gastrointestinal bleeding from ED triage and ROS data	ED	BERT	7144 cases	CDMS
Tahayori et al [25], 2021	Australia	Methodology	Predict patient disposition from ED triage notes	ED	BERT	249,532 ED encounters	CDMS and EWIM
Kim et al [26], 2021	South Korea	Case study	Assign triage severity to simulated cases	ED	BERT	762 cases	CDMS
Wang et al [27], 2021	China	Methodology	Predict diagnosis and appropriate hospital team from medical record	Prehospital	BERT and ClinicalBERT	198,000 patient records	EWIM
McMaster et al [28], 2021	Australia	Methodology	Identify adverse drug events from discharge summaries	Inpatient	BERT (ClinicalBERT and DeBERTa[f])	861 discharge summaries	EWIM
Chen et al [29], 2021	Taiwan	Methodology	Classify electronic health record data into disease presentations	ED	BERT	1,040,989 ED visits and 305,897 NHAMCS[g] samples	EWIM
Drozdov et al [30], 2021	United Kingdom	Methodology	Generate annotations for CXRs[h] to train model to identify COVID-19 cases	ED	BERT (to generate image annotations)	214,042 CXRs	CDMS
Zhang et al [31], 2022	China	Methodology	Classify EMS cases into disease categories	EMS and prehospital	BERT	3500 records	EWIM
Pease et al [32], 2023	United States	Qualitative investigation	Determine the attitudes of clinicians toward using AI[i] in suicide screening	ED	N/A[j]	3 clinicians	CDMS and RET[k]
Chae et al [33], 2023	United States	Methodology	Predict ED visits and hospitalizations for patients with heart failure	Prehospital (home health care)	BERT (BioclinicalBERT)	9362 patients	CDMS and RET

Study	Country	Study type	Purpose	Setting and context	Large language models used	Sample size	Themes
Huang et al [34], 2023	United States	Methodology	Predict nonaccidental trauma	ED	BERT	244,326 trajectories (test) and 2,077,852 trajectories (validation)	CDMS
Chen et al [35], 2023	Taiwan	Methodology	Predict critical outcomes from ED data	ED	BERT (comparator)	171,275 ED visits	CDMS
Smith et al [36], 2023	Australia	Case study	Determine model performance on EM[l] accreditation examination	ED	GPT-3.5 (OpenAI), GPT-4 (OpenAI), Bard-PaLM[m], Bard-PaLM 2, and Bing (Microsoft Corporation)	240 questions	CDMS, RET, and EC[n]
Gupta et al [37], 2023	United States	Case study	Determine the ability of the model to correctly diagnose simulated cases	ED	ChatGPT	20 cases	CDMS, RET, and EC
Abavisani et al [38], 2023	Iran	Commentary	Potential uses of the model in emergency surgery	Emergency surgery	ChatGPT	N/A	CDMS and RET
Rahman et al [39], 2023	United States	Methodology	Identify cases and patterns in unstructured EMS data	EMS and prehospital	BERT (BioBERT and ClinicaBERT)	40,000 EMS narratives	EWIM
Lam and Au [40], 2023	China	Case study	Evaluate model response to lay questions regarding stroke	General public	ChatGPT	3 questions	EC
Bushuven et al [41], 2023	Germany	Case study	Use of the model to advise parents during pediatric emergencies	General public	ChatGPT and GPT-4	22 cases	CDMS, RET, and EC
Ahn [42], 2023	South Korea	Case study	Use of model to provide a lay-person instruction for cardiopulmonary resuscitation	General public	ChatGPT	3 questions	RET and EC
Preiksaitis et al [43], 2023	United States	Commentary	Potential limitations to using models for clinical charting	General medicine	ChatGPT	N/A	EWIM and RET
Barash et al [44], 2023	Israel	Case study	Use of model to aid radiology referral in the ED	ED	GPT-4	40 cases	CDMS and RET
Dahdah et al [45], 2023	United States	Case study	Use of model to triage based on chief complaints	ED	ChatGPT	30 questions	CDMS and RET
Gottlieb et al [46], 2023	United States	Commentary	Discuss advantages and disadvantages of using the model in research	ED and research	ChatGPT	N/A	RET and EC
Babl and Babl [47], 2023	Australia	Case study	Determine the ability of the model to generate a scientific abstract	Research	ChatGPT	1 abstract	RET and EC
Chen et al [48], 2023	China	Methodology	Use the model to study the functioning of web-based self-organizations	Social media	BERT	47,173 users	EWIM
Bradshaw [49], 2023	United States	Case study	Determine the ability of the model to generate discharge instructions	ED	ChatGPT	1 set of discharge instructions	EWIM and EC
Cheng et al [50], 2023	China	Commentary	Potential uses for the model in surgical management	ED	ChatGPT	N/A	CDMS and EWIM

Study	Country	Study type	Purpose	Setting and context	Large language models used	Sample size	Themes
Rao et al [51], 2023	United States	Case study	Test the model performance in several clinical scenarios	General medicine	ChatGPT	36 clinical vignette	EWIM and EC
Brown et al [52], 2023	Jersey	Case report and commentary	Discuss possible model uses in supporting decision-making and clinical care	ED	ChatGPT	1 case	CDMS and EWIM, RET and EC
Bhattaram et al [53], 2023	India	Case study	The ability of the model to triage clinical scenarios	ED	ChatGPT	5 scenarios	CDMS, RET and EC
Webb [54], 2023	United States	Case study	The ability of the model to be used as a communication skill trainer	ED	ChatGPT-3.5	1 case	RET and EC
Hamed et al [55], 2023	Qatar	Case study	The ability of the model to synthesize clinical practice guidelines for diabetic ketoacidosis	General medicine	ChatGPT	3 guidelines	EWIM and RET
Altamimi et al [56], 2023	Saudi Arabia	Case study	The ability of the model to recommend management in snakebites	ED	ChatGPT	9 questions	CDMS and RET
Gebrael et al [57], 2023	United States	Case study	Predict the disposition of patients with metastatic prostate cancer based on ED documentation	ED	ChatGPT-4	56 patients	CDMS, EWIM, and RET
Sarbay et al [58], 2023	Turkey	Case study	Use of the model for patient triage using clinical scenarios	ED	ChatGPT	50 case scenarios	CDMS, EWIM, and RET
Okada et al [59], 2023	Singapore	Commentary	Discuss possible applications for the model in resuscitation	ED or intensive care unit	GPT-3 and GPT-4	N/A	CDMS, EWIM, and RET
Chenais et al [60], 2023	France	Commentary	Describe the landscape of AI-based applications currently in use in EM	ED	BERT and GPT-2	N/A	CDMS, EWIM, and RET

[a]ED: emergency department.

[b]CDMS: clinical decision-making and support.

[c]EWIM: efficiency, workflow, and information management.

[d]BERT: Bidirectional Encoder Representations from Transformers.

[e]EMS: emergency medical service.

[f]DeBERTa: decoding-enhanced Bidirectional Encoder Representations from Transformers with disentangled information.

[g]NHAMCS: National Hospital Ambulatory Medical Care Survey.

[h]CXR: chest x-ray.

[i]AI: artificial intelligence.

[j]N/A: not applicable.

[k]RET: risks, ethics, and transparency.

[l]EM: emergency medicine.

[m]PaLM: Pathways Language Model.

[n]EC: education and communication.

Table 3. Major themes identified, associated subthemes, and representative quotations.

Major theme and subtheme	Representative quotation
Theme 1: clinical decision-making and support	
Prediction	"Machine-learning and natural language processing can be together applied to the ED triage note to predict patient disposition with a high level of accuracy." [25]
Treatment recommendations	"An under-explored use of AI in medicine is predicting and synthesizing patient diagnoses, treatment plans, and outcomes." [51]
Symptom checking and self-triage	"To our knowledge, this is the first work to investigate the capabilities of ChatGPT and GPT-4 on PALS core cases in the hypothetical scenario that laypersons would use the chatbot for support until EMS arrive." [41]
Classification	"In this proof-of-concept study, we demonstrated the process of developing a reliable NER [named-entity recognition] model that could reliably identify clinical entities from unlabeled paramedic free text reports." [22]
Triage	"...this preliminary study showed the potential of developing an automatic classification system that directly classifies the KTAS [triage] level and symptoms from the conversations between patients and clinicians." [26]
Screening	"We showed that PABLO, a pretrained, domain-adapted outcome forecasting model, can be used to predict both first and recurrent instances of NAT [non-accidental trauma]." [34]
Differential diagnosis building	"These results suggest that ChatGPT has a high level of accuracy in predicting top differential diagnoses in simulated medical cases." [37]
Decision support	"...ChatGPT-4 demonstrates encouraging results as a support tool in the ED. LLMs such as ChatGPT-4 can facilitate appropriate imaging examination selection and improve radiology referral quality." [44]
Clinical augmentation	"AI can serve as an adjunct in clinical decision making throughout the entire clinical workflow, from triage to diagnosis to management." [51]
Theme 2: efficiency, workflow, and information management	
Unstructured data extraction	"The proposed model will provide a method to further extract the unstructured free-text portions in EHRs to obtain an abundance of health data. As we enter the forefront of the artificial intelligence era, NLP deep-learning models are well under development. In our model, all medical free-text data can be transformed into meaningful embeddings, which will enhance medical studies and strengthen doctors' capabilities." [20]
Charting efficiency	"While notes have become more structured and burdensome, the field of data science has rapidly advanced. With such powerful tools available, it seems reasonable to explore their use to automate seemingly mundane tasks such as writing clinical notes. Generative AI models like ChatGPT could be developed to populate notes for patients based on massive amounts of data contained in current EHRs." [43]
Summarization or synthesis	"Although ChatGPT demonstrates the potential for the synthesis of clinical guidelines, the presence of multiple recurrent errors and inconsistencies underscores the need for expert human intervention and validation." [55]
Pattern identification	"This embedding system can be used as a disease retrieval model, which encodes queries and finds the most relevant patients and diseases. In the retrieval demonstration, the query subject was a 53-year-old female patient who suffered from abdominal pain in the upper right quarter to right flanks for 3 days and noticed dizziness and tarry stool on the day of the interview. Through the retrieval, we obtained the five most similar patients with similar symptoms that were possibly related to different diseases." [29]
Workflow efficiency	"Integration of LLMs with existing EHR (with appropriate regulations) could facilitate improved patient outcomes and workflow efficiency." [51]
Theme 3: risks, ethics, and transparency	
Oversight	"Generally speaking, the Ethics Guideline for Trustworthy AI suggested seven key requirements including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, nondiscrimination and fairness, environmental and societal well-being, and accountability." [59]
Fairness	"[Use of LLMs] could also increase equity by assisting researchers with disabilities such as dyslexia." [46]
Ethical and legal responsibilities	"Legal and ethical implications are associated with using AI in clinical practice, particularly regarding privacy and informed consent issues." [52]

Major theme and subtheme	Representative quotation
Reliance on input data	"...data quality can affect the performance of LLMs and NLP techniques applied to the task of extracting and summarizing clinical guidelines." [55]
Overreliance	"Overreliance on AI systems and the assumption that they are infallible or less fallible than human judgment–automation bias–can lead to errors." [52]
Explainability and transparency	"Creating a clinician-interpretable risk prediction model is essential for clinical adoption and implementation of models because it builds trust in decisionmakers, enables error identification and correction in the model, and facilitates integration into clinical workflows." [33]
Bias propagation	"A risk of bias is possible if the initial training data is not representative of the study population. There is a possibility of compounding of bias and error, leading to incorrect assessment." [53]
Human bias reduction	"AI tools can offer a near real-time interpretation of medical imaging and clinical decision support and may identify latent patterns that may not be evident to clinicians. While humans are prone to cognitive biases, such as prejudice or fatigue, which can hinder their decision-making process, AI can mitigate these biases and improve accuracy in patient care." [52]
Accuracy	"LLMs may not be exposed to the broader range of literature (particularly if studies are located behind paywalls), which may limit the comprehensiveness or accuracy of the data." [46]
Theme 4: education and communication	
Clinician education	"While LLM performance in medical examinations may initially seem to be little more than a novelty, their ability to generate coherent and well-explained content hints at other potential uses. As a medical education tool they could potentially help generate practice questions, design mock examinations or provide additional explanations for complex concepts." [36]
Communication	"Although in its infancy, AI chatbot use has the potential to disrupt how we teach medical students and graduate medical residents communication skills in outpatient and hospital settings." [54]
Content generation	"ChatGPT or similar programmes, with careful review of the product by authors, may become a valuable scientific writing tool." [47]
Research assistance	"Conversational AI has some clear benefits and disadvantages. As the technology further evolves, it is incumbent on the scientific community to determine how best to incorporate LLMs into the research and publication process with attention to scientific integrity, adherence to ethical principles, and existing copyright laws." [46]

Theme 1: Clinical Decision-Making and Support

The first theme we identified is clinical decision-making and support. LLMs have been used or proposed for applications such as providing advice to the public before arrival; aiding in triage as patients arrive at the ED; or augmenting the activities of physicians as they provide care, either through supporting diagnostics or predicting patient resource use.

Several applications focused on advising the public and aiding in symptom checking, self-triage, and occasionally advising first-aid before the arrival of emergency medical services. These included counseling parents during potential pediatric emergencies, recognizing stroke, or providing advice during potential cardiac arrests [40-42]. Wang et al [27] proposed a model that could potentially help patients navigate the complexities of the health care system in China and present to the correct medical setting for the care they need.

Furthermore, LLMs have the potential to efficiently screen patients for important outcomes, such as pediatric patients at risk for nonaccidental trauma, suicide risk, or COVID-19 infection [30,32,34]. These can be implemented based on data in the medical record or as clinical data are obtained in real time.

Early identification of patient risks could help physicians more rapidly identify important diagnoses. Several studies discussed implementations of LLMs that work in conjunction with physicians while caring for patients in the ED [50,51]. Brown et al [52] discuss the potential role of these models in overcoming cognitive biases and reducing errors. These models could be used in developing a differential diagnosis, recommending imaging studies, providing treatment recommendations, or interpreting clinical guidelines [37,44,55,56].

Several studies centered on predicting outcomes such as presentation to the ED, hospitalization, intensive care unit admission, or in-hospital cardiac arrest [25,33,35,57]. Applications of LLMs in the triage process could potentially identify patients who require immediate attention or patients at a high risk of certain diagnoses, such as gastrointestinal bleeding [24,26,53,58,60].

Theme 2: Efficiency, Workflow, and Information Management

The second theme identified is information management, workflow, and efficiency. LLMs show great promise in increasing the usability of data available in the EHR. Interactions with the EHR take up a substantial amount of physician time, and it is often difficult to identify crucial information during critical times [43]. LLMs could serve a variety of information management functions. They could be used to perform audits for quality improvement purposes, identify potential adverse events such as drug interactions, anticipate and monitor public health emergencies, and assist with information entry during

the clinical encounter [19,20,22,23,28,31,39,43,49]. LLMs developed and trained on data from the ED could quickly identify similar patient presentations, recognize patterns, and extract important information from unstructured text [18,20,21,60].

Some authors suggest that LLMs can enhance care throughout the entire EM encounter [30,50-52]. LLMs could potentially be used as digital adjuncts for clinical decision-making because they could generate differentials, predict final diagnoses, offer interpretations of imaging studies, and suggest treatment plans [30,51,52,61]. They may mitigate human cognitive biases and address human factors (eg, time constraints, frequent task switching, high cognitive load, constant interruptions, and decision fatigue) that predispose emergency physicians to error [52].

The flexibility and versatility of the LLMs offer particular benefits to EM practice. The diverse ways in which these models can aid throughout the entire clinical workflow could help physicians process large quantities of complex clinical data, mitigate cognitive biases, and deliver relevant information in a comprehensible format [30,51,52,61]. By streamlining these burdensome tasks, LLMs could help improve the efficiency of care for the high volume of patients the physicians routinely see in the ED.

Theme 3: Risks, Transparency, and Ethics

Despite the potential for advancement and improvement in the care that EM physicians can provide through the inclusion of LLMs in practice, several issues limit their implementation into practice at this time.

The most often discussed risk, mentioned in 11 (26%) of the 43 papers, is the reliability of model responses and the potential for erroneous results [20,21,28-30,44,51,53,55,56,59]. These output errors often result from inaccuracies in the training data, which are most commonly gathered from the internet and unvetted for reliability. Sources of inaccurate responses may be identified by examining the training material, but other errors due to data noise, mislabeling, or outdated information may be harder to detect [21,28,30,56]. Similarly, biases in training data can be propagated to the model, leading to inaccurate or discriminatory results [51,53,57,60,62]. In medical applications, the consequences of the errors can be significant, and even small errors could lead to adverse outcomes [51].

Understanding and mitigating errors in LLMs is challenging due to issues with transparency and reproducibility of model outputs [52-54,59,62]. Better understanding among clinicians of the algorithms and statistical methods used by LLMs is a suggested method to ensure cautious use [52]. Concentrating on making models more explainable or transparent is another potential approach [62]. However, the degree to which this will be feasible, given the complexity of these models, remains to be determined.

Patient and data privacy is another clearly articulated risk of using these models in the clinical environment [35,52,53]. There are some proposed methodologies using unsupervised methods that can train the models with limited access to sensitive information; however, these require further exploration [35]. Patient attitudes and willingness to allow models access to their health information for training and how to address disclosure of this use have not been extensively discussed. Finally, the legal and ethical implications of using LLM output to guide patient care is an often-mentioned concern [52,53,59]. How the responsibility for patient care decisions is distributed if LLMs are used to guide clinical decisions is yet to be determined.

Theme 4: Education and Communication

LLMs offer several opportunities for education and communication. First, several papers noted that the successful integration of LLMs into clinical practice will require physicians to understand the underlying algorithms and statistical methods used by these models [52,59]. There is a need for dedicated educational programs on AI in medicine at all levels of medical education to ensure that the solutions developed align with the clinical environment and address the unique challenges of working with clinical data [34,51,63].

In terms of clinical education, several studies have demonstrated reasonable performance of LLMs on standardized tests in medicine, which could indicate the potential for these models to develop study materials [36]. In addition, these models may be able to help physicians communicate with and educate the patients. Dahdah et al [45] used ChatGPT to answer several common medical questions in easy-to-understand language, suggesting the ability to enhance physician responses to patient queries. Webb [54] demonstrated the use of ChatGPT to simulate patient conversation and provide feedback to a physician learning how to break bad news.

Patient education may be facilitated via these models without physician input as well. As discussed in the previous sections, several authors described applications designed to educate patients during emergencies before they arrived in the ED [27,40-42]. Finally, LLMs could be used to aid in knowledge dissemination. Gottleib et al [46] and Babl and Babl [47] describe potential applications for LLMs in research and scientific writing. They highlight potential benefits to individuals who struggle with English or have challenges with writing or knowledge synthesis. In addition, models may be used to translate scientific papers more rapidly. However, the use of these models to generate scientific papers raises concerns regarding the potential for academic dishonesty [46,47].

Discussion

Principal Findings

Our review aligns with the growing body of literature emphasizing the great potential for AI in EM, particularly in areas such as time-sensitive decision-making and managing high-volume data [2-5,60]. However, our focus on LLMs and their unique capabilities extends the current understanding of AI applications in EM. Although several specific applications and limitations have been reported and suggested in the literature, our analysis identified 4 major areas of focus for LLMs in EM: clinical decision support, workflow efficiency, risks, ethics, and education. We propose these topics as a framework for understanding emerging implementations of LLMs and as a guide to inform future areas of investigation.

At their core, LLMs and their associated natural language processing techniques offer a way to organize and engage with vast amounts of unstructured text data. Depending on how they are trained and used, they can be operationalized to make predictions or identify patterns, which gives rise to most of our identified applications. Most commercially available LLMs, such as ChatGPT, are trained on massive volumes of text gathered from the internet and then optimized for conversational interaction [64]. This ability to access a breadth of general knowledge and the resulting wide applicability have contributed to the increased use of LLMs by professionals and the public across a variety of fields [65]. As these models become more ubiquitous, there is potential for their use across the care continuum. They could not only support clinical care but also provide an opportunity to offer advice to the public regarding medical concerns. Several papers (3/34, 9%) in our review identified the feasibility of using LLMs to provide first-aid instructions and offer decision support to potential patients seeking care [40-42].

Preliminary work suggests that dedicated training can enhance the ability of these models to make triage recommendations, but prospective implementation has not been tested [27]. LLMs could certainly aid patients in self-triage or with basic medical questions; nevertheless, how this can be effectively and safely implemented needs further exploration, especially with concerns regarding the accuracy of outputs. Possibilities to improve outputs include additional dedicated training of the models to align with the medical and emergency settings to improve their reliability and accuracy. These context-specific models could be equipped with information on the local health care system to help patients identify available resources, schedule appointments, or activate emergency medical services.

In the ED, LLMs could increase workflow efficiency by rapidly synthesizing relevant information from a patient's medical record, structuring and categorizing chief complaint data, and assigning an emergency severity index level [18,21,26,45,53,58]. In addition, quickly accessing data from the medical record could improve the efficiency and thoroughness of chart review. A model's ability to identify subtle patterns in data could offer additional diagnostic support by recommending or interpreting laboratory and imaging studies [30,51,52,61]. By facilitating tasks such as information retrieval and synthesis, LLMs could reduce this burden for clinicians and minimize errors due to buried or disorganized data, potentially contributing to workflow efficiency. Furthermore, they may counteract human cognitive biases and fatigue when used to support clinical decisions [52]. Although some studies have demonstrated reasonable accuracy on focused use cases, further validation of any of these applications across diverse settings and patient populations is required. Thoughtful integration of LLMs has the potential to revolutionize EM by providing clinical decision support, improving situational awareness, and increasing productivity.

However, barriers to seamless implementation exist. As noted by several authors, erroneous outputs remain a concern, given the dependence on training data [28-30,35,51,53,55,56,59]. Information surrounding the most publicly available LLMs today is obscured across three important layers: (1) the underlying training data used—commonly reported to be publicly available data on the internet and from third-party licensed data sets, (2) the underlying architecture of the model—whose exact mechanisms are not always easy to discern, and (3) the intricacies of human-led fine-tuning—often done at the end of development to provide guardrails for output. These layers of obscurity make it difficult to troubleshoot the cause of any single erroneous output.

Regarding privacy and data rights, it is imperative to discuss and implement privacy-preserving methods for patient data. The use of techniques such as data anonymization, differential privacy, and federated learning are instrumental in safeguarding patient information. Data anonymization involves removing or modifying personal identifiers to prevent the association of data with individual patients. Differential privacy introduces randomness into the data or queries to ensure individual data points cannot be isolated [66]. Federated learning enables models to be trained against multiple decentralized devices or servers holding local data samples without exchanging them, thus enhancing privacy [67]. The specific ways in which LLMs will interface with other hospital information systems, such as the EHR, need further exploration, and careful integration is critical to address privacy concerns, especially given the sensitive nature of health care data.

Moreover, the ongoing discussions about the information used in these models underscore the need for continuous scrutiny [52,53,59]. In addition to privacy, the legal and ethical implications of AI-assisted health care require further exploration to establish robust oversight and accountability structures. Without a commitment to explainability and transparency, the use of *black box* LLMs may encounter resistance from clinicians.

Our review reveals several opportunities for future exploration and research. Perhaps the most important is effectively identifying problems that are best solved using LLMs in EM. Our review outlines several immediate areas of potential exploration, including improved communication, translation, and summarization of highly detailed and domain-specific knowledge for providers and patients, but further exploration and prospective validation of specific use cases is required. We expect the potential use cases in EM to grow as LLMs become increasingly complex and develop emergent properties–actions that are not explicitly programmed or anticipated. To bridge the *AI chasm* between innovations in the research realm and widespread adoption, these applications should be identified with significant input from providers in the clinical space who can uniquely identify areas of potential benefit. To accomplish this, a better understanding of the abilities and limitations of LLMs among physicians is needed to optimize their best use and ensure they are effectively implemented, and AI literacy is increasingly described as an essential competency for physicians [68]. We encourage the development of curricula and training programs designed for emergency physicians.

Given the black-box nature of LLMs, standardized frameworks and metrics for evaluation that are specific to health care use cases are needed to evaluate their performance and implementation effectively. These frameworks should encompass an understanding of both the technical capabilities

and constraints of a model, along with the human interaction aspects that affect its use. A crucial part of this assessment involves comparing the performance of LLMs to human proficiency, determining whether the objective is to replace or enhance tasks currently carried out by health care professionals. Thorough testing of models in real time, real-world scenarios is imperative before their deployment. The selection of patient- or provider-focused outcomes is essential, and the effectiveness of models should not be evaluated in isolation. Instead, it is crucial to assess the combined performance of the provider and AI system to ensure that models are effective and practical in real-world settings. Implementing and validating solutions should occur across diverse populations and care environments, with particular focus on cohorts underrepresented in the training data to mitigate potential harm from model biases [69]. Provider perspectives are essential, but equally important are patient perspectives about the use of LLMs in medicine. Impacts on physician-patient communication, patient concerns surrounding privacy, and attitudes toward AI-generated recommendations must be further explored. Collaboration between all relevant stakeholders who develop or will be impacted by LLMs for clinical medicine is essential for developing models that can be used effectively, equitably, and safely.

Limitations

This scoping review has some limitations worth noting. First, we restricted our search to papers published after 2018, when LLMs first emerged. While this captures the current era of LLMs, earlier works relevant to natural language processing in EM may have been overlooked. In addition, despite searching 4 databases and consulting a medical librarian on the search strategy, some pertinent studies may have been missed, and given the rapidly evolving nature of this research area, there are certainly more studies that have emerged since our literature search [70]. However, our review establishes an initial foundation that can be built upon as the field continues to grow. Finally, in an effort to be maximally inclusive in our review, we did not include or exclude papers based on the quality of their evidence. Similarly, we did not make any quality determinations of our included studies. High-quality studies are required to make any determination regarding the efficacy of LLMs for the applications we described, and our review hopefully provides a framework to design these investigations.

Conclusions

This review underscores the transformative potential of LLMs in enhancing the delivery of emergency care. By leveraging their ability to process vast amounts of data rapidly, LLMs offer unprecedented opportunities to improve decision-making speed and accuracy, a critical component in the high-stakes, fast-paced EM environment. From the identified themes, it is evident that LLMs have the potential to revolutionize various aspects of emergency care, highlighting their versatility and the breadth of their applicability.

From the theme of clinical decision-making and support, LLMs can augment the diagnostic process, support differential diagnosis, and aid in the efficient allocation of resources. In the domain of efficiency, workflow, and information management, LLMs have shown promise in enhancing operational efficiencies, reducing the cognitive load on clinicians, and streamlining patient care processes. Regarding risks, ethics, and transparency, the review illuminates the need for meticulous attention to the accuracy, bias, and ethical considerations inherent in deploying LLMs in a clinical setting. Finally, in the realm of education and communication, LLMs' potential to facilitate learning and improve patient and provider communication signifies a paradigm shift in medical education and engagement.

The most urgent research need identified in this review is the development of robust, evidence-based frameworks for evaluating the clinical efficacy of LLMs in EM; addressing ethical concerns; ensuring data privacy; and mitigating potential biases in model outputs. There is a critical need for prospective studies that validate the utility of LLMs in real-world emergency care settings and explore the optimization of these models for specific clinical tasks. Furthermore, research should focus on understanding the best practices for integrating LLMs into the existing health care workflows without disrupting the clinician-patient relationship.

The successful integration of LLMs into EM necessitates a multidisciplinary approach involving clinicians, computer scientists, ethicists, patients, and policy makers. Collaborative efforts are essential to navigate the challenges of implementing AI technologies in health care, ensuring LLMs complement the clinical judgment of EM professionals and align with the overarching goal of improving patient care. The judicious application of LLMs has the potential to fundamentally redefine much of EM practice, ushering in a future where care is more accurate, efficient, and responsive to the needs of patients. Furthermore, by reducing the many burdens that currently encumber clinicians, these technologies hold the promise of restoring and deepening the invaluable human connections between physicians and their patients.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature review search strategy.
[DOCX File , 14 KB - medinform_v12i1e53787_app1.docx]

Multimedia Appendix 2

PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 630 KB - medinform_v12i1e53787_app2.pdf]

References

1. Petrino R, Riesgo LG, Yilmaz B. Burnout in emergency medicine professionals after 2 years of the COVID-19 pandemic: a threat to the healthcare system? Eur J Emerg Med 2022 Aug 01;29(4):279-284 [FREE Full text] [doi: 10.1097/MEJ.0000000000000952] [Medline: 35620812]
2. Piliuk K, Tomforde S. Artificial intelligence in emergency medicine. A systematic literature review. Int J Med Inform 2023 Dec;180:105274 [FREE Full text] [doi: 10.1016/j.ijmedinf.2023.105274] [Medline: 37944275]
3. Kirubarajan A, Taher A, Khan S, Masood S. Artificial intelligence in emergency medicine: a scoping review. J Am Coll Emerg Physicians Open 2020 Nov 07;1(6):1691-1702 [FREE Full text] [doi: 10.1002/emp2.12277] [Medline: 33392578]
4. Masoumian Hosseini M, Masoumian Hosseini ST, Qayumi K, Ahmady S, Koohestani HR. The aspects of running artificial intelligence in emergency care; a scoping review. Arch Acad Emerg Med 2023 May 11;11(1):e38 [FREE Full text] [doi: 10.22037/aaem.v11i1.1974] [Medline: 37215232]
5. Mueller B, Kinoshita T, Peebles A, Graber MA, Lee S. Artificial intelligence and machine learning in emergency medicine: a narrative review. Acute Med Surg 2022 Mar 1;9(1):e740 [FREE Full text] [doi: 10.1002/ams2.740] [Medline: 35251669]
6. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
7. 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]
8. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb 23;8(1):19-32. [doi: 10.1080/1364557032000119616]
9. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
10. Preiksaitis C. Protocol for a scoping review of the application of large language models in emergency medicine. OSF Home. 2023 Oct 19. URL: https://osf.io/tdghu/ [accessed 2024-04-28]
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online October 11, 2018 2024(https://arxiv.org/abs/1810.04805) [FREE Full text] [doi: 10.5260/chara.21.2.8]
12. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
13. Brown TB, 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 2024. [doi: 10.5860/choice.189890]
14. Schreiner M. GPT-4 architecture, datasets, costs and more leaked. The Decoder. 2023 Jul 11. URL: https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/ [accessed 2023-10-12]
15. Narang S, Chowdhery A. Pathways language model (PaLM): scaling to 540 billion parameters for breakthrough performance. Google Research. 2022 Apr 04. URL: https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html [accessed 2023-10-12]
16. AllenNLP - ELMo. Allen Institute for Artificial Intelligence. URL: https://allenai.org/allennlp/software/elmo [accessed 2023-10-12]
17. Devlin J, Chang MW. Open sourcing BERT: state-of-the-art pre-training for natural language processing. Google Research. URL: https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html [accessed 2023-10-12]
18. Xu B, Gil-Jardiné C, Thiessard F, Tellier E, Avalos 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 2024.
19. Wang T, Lu K, Chow KP, Zhu Q. COVID-19 sensing: negative sentiment analysis on social media in China via BERT model. IEEE Access 2020 Jul 28;8:138162-138169. [doi: 10.1109/access.2020.3012595]
20. 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]
21. Chang D, Hong WS, Taylor RA. Generating contextual embeddings for emergency department chief complaints. JAMIA Open 2020 Jul 15;3(2):160-166 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa022] [Medline: 32734154]
22. Wang H, Yeung WL, Ng QX, Tung A, Tay JA, Ryanputra D, et al. A weakly-supervised named entity recognition machine learning approach for emergency medical services clinical audit. Int J Environ Res Public Health 2021 Jul 22;18(15):7776 [FREE Full text] [doi: 10.3390/ijerph18157776] [Medline: 34360065]
23. Gil-Jardiné C, Chenais G, Pradeau C, Tentillier E, Revel P, Combes X, et al. Trends in reasons for emergency calls during the COVID-19 crisis in the department of Gironde, France using artificial neural network for natural language classification.

Scand J Trauma Resusc Emerg Med 2021 Mar 31;29(1):55 [FREE Full text] [doi: 10.1186/s13049-021-00862-w] [Medline: 33789721]
24. Shung D, Tsay C, Laine L, Chang D, Li F, Thomas P, et al. Early identification of patients with acute gastrointestinal bleeding using natural language processing and decision rules. J Gastroenterol Hepatol 2021 Jun;36(6):1590-1597. [doi: 10.1111/jgh.15313] [Medline: 33105045]
25. Tahayori B, Chini-Foroush N, Akhlaghi H. Advanced natural language processing technique to predict patient disposition based on emergency triage notes. Emerg Med Australas 2021 Jun;33(3):480-484. [doi: 10.1111/1742-6723.13656] [Medline: 33043570]
26. Kim D, Oh J, Im H, Yoon M, Park J, Lee J. Automatic classification of the Korean triage acuity scale in simulated emergency rooms using speech recognition and natural language processing: a proof of concept study. J Korean Med Sci 2021 Jul 12;36(27):e175 [FREE Full text] [doi: 10.3346/jkms.2021.36.e175] [Medline: 34254471]
27. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using Chinese medical BERT. J Cloud Comput 2021 Jan 15;10:4. [doi: 10.1186/s13677-020-00218-2]
28. McMaster C, Chan J, Liew DF, Su E, Frauman AG, Chapman WW, et al. Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. J Biomed Inform 2023 Jan;137:104265 [FREE Full text] [doi: 10.1016/j.jbi.2022.104265] [Medline: 36464227]
29. Chen YP, Lo YH, Lai F, Huang CH. Disease concept-embedding based on the self-supervised method for medical information extraction from electronic health records and disease retrieval: algorithm development and validation study. J Med Internet Res 2021 Jan 27;23(1):e25113 [FREE Full text] [doi: 10.2196/25113] [Medline: 33502324]
30. Drozdov I, Szubert B, Reda E, Makary P, Forbes D, Chang SL, et al. Development and prospective validation of COVID-19 chest X-ray screening model for patients attending emergency departments. Sci Rep 2021 Oct 14;11(1):20384 [FREE Full text] [doi: 10.1038/s41598-021-99986-3] [Medline: 34650190]
31. Zhang X, Zhang H, Sheng L, Tian F. DL-PER: deep learning model for Chinese prehospital emergency record classification. IEEE Access 2022 Jun 03;10:64638-64649. [doi: 10.1109/ACCESS.2022.3179685]
32. Pease JL, Thompson D, Wright-Berryman J, Campbell M. User feedback on the use of a natural language processing application to screen for suicide risk in the emergency department. J Behav Health Serv Res 2023 Oct 03;50(4):548-554 [FREE Full text] [doi: 10.1007/s11414-023-09831-w] [Medline: 36737559]
33. Chae S, Davoudi A, Song J, Evans L, Hobensack M, Bowles KH, et al. Predicting emergency department visits and hospitalizations for patients with heart failure in home healthcare using a time series risk model. J Am Med Inform Assoc 2023 Sep 25;30(10):1622-1633. [doi: 10.1093/jamia/ocad129] [Medline: 37433577]
34. Huang D, Cogill S, Hsia RY, Yang S, Kim D. Development and external validation of a pretrained deep learning model for the prediction of non-accidental trauma. NPJ Digit Med 2023 Jul 19;6(1):131 [FREE Full text] [doi: 10.1038/s41746-023-00875-y] [Medline: 37468526]
35. Chen MC, Huang TY, Chen TY, Boonyarat P, Chang YC. Clinical narrative-aware deep neural network for emergency department critical outcome prediction. J Biomed Inform 2023 Feb;138:104284 [FREE Full text] [doi: 10.1016/j.jbi.2023.104284] [Medline: 36632861]
36. Smith J, Choi PM, Buntine P. Will code one day run a code? Performance of language models on ACEM primary examinations and implications. Emerg Med Australas 2023 Oct;35(5):876-878. [doi: 10.1111/1742-6723.14280] [Medline: 37414729]
37. Gupta P, Nayak R, Alazzeh M. The accuracy of medical diagnoses in emergency medicine by modern artificial intelligence. Acad Emerg Med 2023;30(Suppl 1):395 [FREE Full text] [doi: 10.1111/acem.14718]
38. Abavisani M, Dadgar F, Keikha M. A commentary on emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor's right-hand man. Int J Surg 2023 Oct 01;109(10):3195-3196 [FREE Full text] [doi: 10.1097/JS9.0000000000000561] [Medline: 37318859]
39. Rahman MA, Preum SM, Williams RD, Alemzadeh H, Stankovic J. EMS-BERT: a pre-trained language representation model for the emergency medical services (EMS) domain. In: Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies. 2023 Presented at: CHASE '23; June 21-23, 2023; Orlando, FL. [doi: 10.1145/3580252.3586978]
40. Lam WY, Au SC. Stroke care in the ChatGPT era: potential use in early symptom recognition. J Acute Dis 2023 Jun;12(3):129-130. [doi: 10.4103/2221-6189.379278]
41. Bushuven S, Bentele M, Bentele S, Gerber B, Bansbach J, Ganter J, et al. “ChatGPT, can you help me save my child’s life?” - diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital Basic Life Support and Paediatric Advanced Life Support cases – an in-silico analysis. Research Square. Preprint posted online May 12, 2023 2024 [FREE Full text] [doi: 10.21203/rs.3.rs-2910261/v1]
42. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation 2023 Apr;185:109729. [doi: 10.1016/j.resuscitation.2023.109729] [Medline: 36773836]
43. Preiksaitis C, Sinsky CA, Rose C. ChatGPT is not the solution to physicians' documentation burden. Nat Med 2023 Jun;29(6):1296-1297. [doi: 10.1038/s41591-023-02341-4] [Medline: 37169865]

44. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 assistance in optimizing emergency department radiology referrals and imaging selection. J Am Coll Radiol 2023 Oct;20(10):998-1003. [doi: 10.1016/j.jacr.2023.06.009] [Medline: 37423350]
45. Dahdah JE, Kassab J, Helou MC, Gaballa A, Sayles S3, Phelan MP. ChatGPT: a valuable tool for emergency medical assistance. Ann Emerg Med 2023 Sep;82(3):411-413. [doi: 10.1016/j.annemergmed.2023.04.027] [Medline: 37330721]
46. Gottlieb M, Kline JA, Schneider AJ, Coates WC. ChatGPT and conversational artificial intelligence: friend, foe, or future of research? Am J Emerg Med 2023 Aug;70:81-83. [doi: 10.1016/j.ajem.2023.05.018] [Medline: 37229893]
47. Babl FE, Babl MP. Generative artificial intelligence: can ChatGPT write a quality abstract? Emerg Med Australas 2023 Oct;35(5):809-811 [FREE Full text] [doi: 10.1111/1742-6723.14233] [Medline: 37142327]
48. Chen J, Liu Q, Liu X, Wang Y, Nie H, Xie X. Exploring the functioning of online self-organizations during public health emergencies: patterns and mechanism. Int J Environ Res Public Health 2023 Feb 23;20(5):4012 [FREE Full text] [doi: 10.3390/ijerph20054012] [Medline: 36901022]
49. Bradshaw JC. The ChatGPT era: artificial intelligence in emergency medicine. Ann Emerg Med 2023 Jun;81(6):764-765. [doi: 10.1016/j.annemergmed.2023.01.022] [Medline: 37210166]
50. Cheng K, Li Z, Guo Q, Sun Z, Wu H, Li C. Emergency surgery in the era of artificial intelligence: ChatGPT could be the doctor's right-hand man. Int J Surg 2023 Jun 01;109(6):1816-1818 [FREE Full text] [doi: 10.1097/JS9.0000000000000410] [Medline: 37074733]
51. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv. Preprint posted online February 26, 2023 2023 Feb 26 [FREE Full text] [doi: 10.1101/2023.02.21.23285886] [Medline: 36865204]
52. Brown C, Nazeer R, Gibbs A, Le Page P, Mitchell AR. Breaking bias: the role of artificial intelligence in improving clinical decision-making. Cureus 2023 Mar 20;15(3):e36415 [FREE Full text] [doi: 10.7759/cureus.36415] [Medline: 37090406]
53. Bhattaram S, Shinde VS, Khumujam PP. ChatGPT: the next-gen tool for triaging? Am J Emerg Med 2023 Jul;69:215-217. [doi: 10.1016/j.ajem.2023.03.027] [Medline: 37024324]
54. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus 2023 May 09;15(5):e38755 [FREE Full text] [doi: 10.7759/cureus.38755] [Medline: 37303324]
55. Hamed E, Eid A, Alberry M. Exploring ChatGPT's potential in facilitating adaptation of clinical guidelines: a case study of diabetic ketoacidosis guidelines. Cureus 2023 May 09;15(5):e38784 [FREE Full text] [doi: 10.7759/cureus.38784] [Medline: 37303347]
56. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. Cureus 2023 Jun 13;15(6):e40351 [FREE Full text] [doi: 10.7759/cureus.40351] [Medline: 37456381]
57. Gebrael G, Sahu KK, Chigarira B, Tripathi N, Mathew Thomas V, Sayegh N, et al. Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. Cancers (Basel) 2023 Jul 22;15(14):3717 [FREE Full text] [doi: 10.3390/cancers15143717] [Medline: 37509379]
58. Sarbay İ, Berikol G, Özturan İ. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. Turk J Emerg Med 2023 Jun 26;23(3):156-161 [FREE Full text] [doi: 10.4103/tjem.tjem_79_23] [Medline: 37529789]
59. Okada Y, Mertens M, Liu N, Lam SS, Ong ME. AI and machine learning in resuscitation: ongoing research, new concepts, and key challenges. Resusc Plus 2023 Jul 28;15:100435 [FREE Full text] [doi: 10.1016/j.resplu.2023.100435] [Medline: 37547540]
60. Chenais G, Lagarde E, Gil-Jardiné C. Artificial intelligence in emergency medicine: viewpoint of current applications and foreseeable opportunities and challenges. J Med Internet Res 2023 May 23;25:e40031 [FREE Full text] [doi: 10.2196/40031] [Medline: 36972306]
61. Chen HL, Chen HH. Have you chatted today? - medical education surfing with artificial intelligence. J Med Educ 2023 Mar;27(1):1-4 [FREE Full text]
62. Fanconi C, van Buchem M, Hernandez-Boussard T. Natural language processing methods to identify oncology patients at high risk for acute care with clinical notes. AMIA Jt Summits Transl Sci Proc 2023 Jun 16;2023:138-147 [FREE Full text] [Medline: 37350895]
63. Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ 2023 Oct 20;9:e48785 [FREE Full text] [doi: 10.2196/48785] [Medline: 37862079]
64. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-10-06]
65. 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 2023-10-06]
66. Ziller A, Usynin D, Braren R, Makowski M, Rueckert D, Kaissis G. Medical imaging deep learning with differential privacy. Sci Rep 2021 Jun 29;11(1):13524 [FREE Full text] [doi: 10.1038/s41598-021-93030-0] [Medline: 34188157]
67. 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 Sep 14;3:119 [FREE Full text] [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]

68. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024 Jan 01;99(1):22-27. [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
69. Rose C, Barber R, Preiksaitis C, Kim I, Mishra N, Kayser K, et al. A conference (missingness in action) to address missingness in data and AI in health care: qualitative thematic analysis. J Med Internet Res 2023 Nov 23;25:e49314 [FREE Full text] [doi: 10.2196/49314] [Medline: 37995113]
70. Chenais G, Gil-Jardiné C, Touchais H, Avalos Fernandez M, Contrand B, Tellier E, et al. Deep learning transformer models for building a comprehensive and real-time trauma observatory: development and validation study. JMIR AI 2023 Jan 12;2:e40843. [doi: 10.2196/40843]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
ED: emergency department
EHR: electronic health record
EM: emergency medicine
LLM: large language model
MeSH: Medical Subject Headings
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by A Castonguay; submitted 19.10.23; peer-reviewed by L Zhu, C Gil-Jardiné, MO Khursheed; comments to author 13.12.23; revised version received 20.12.23; accepted 05.04.24; published 10.05.24.

Please cite as:
Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, Ribeira R, Rose C
The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review
JMIR Med Inform 2024;12:e53787
URL: https://medinform.jmir.org/2024/1/e53787
doi:10.2196/53787
PMID:38728687

CHDmap: One Step Further Toward Integrating Medicine-Based Evidence Into Practice

Jef Van den Eynde, MD

Department of Cardiovascular Sciences, KU Leuven, Leuven, Belgium

Corresponding Author:
Jef Van den Eynde, MD

Related Article:

https://medinform.jmir.org/2024/1/e49138

Abstract

Evidence-based medicine, rooted in randomized controlled trials, offers treatment estimates for the average patient but struggles to guide individualized care. This challenge is amplified in complex conditions like congenital heart disease due to disease variability and limited trial applicability. To address this, medicine-based evidence was proposed to synthesize information for personalized care. A recent article introduced a patient similarity network, CHDmap, which represents a promising technical rendition of the medicine-based evidence concept. Leveraging comprehensive clinical and echocardiographic data, CHDmap creates an interactive patient map representing individuals with similar attributes. Using a k-nearest neighbor algorithm, CHDmap interactively identifies closely resembling patient groups based on specific characteristics. These approximate matches form the foundation for predictive analyses, including outcomes like hospital length of stay and complications. A key finding is the tool's dual capacity: not only did it corroborate clinical intuition in many scenarios, but in specific instances, it prompted a reevaluation of cases, culminating in an enhancement of overall performance across various classification tasks. While an important first step, future versions of CHDmap may aim to expand mapping complexity, increase data granularity, consider long-term outcomes, allow for treatment comparisons, and implement artificial intelligence–driven weighting of various input variables. Successful implementation of CHDmap and similar tools will require training for practitioners, robust data infrastructure, and interdisciplinary collaboration. Patient similarity networks may become valuable in multidisciplinary discussions, complementing clinicians' expertise. The symbiotic approach bridges evidence, experience, and real-life care, enabling iterative learning for future physicians.

(JMIR Med Inform 2024;12:e52343) doi:10.2196/52343

KEYWORDS

artificial intelligence; clinical practice; congenital heart disease; decision-making; evidence-based medicine; machine learning; medicine-based evidence; patient similarity networks; precision medicine; randomized controlled trials

Evidence-based medicine (EBM), built on the foundations of randomized controlled trials (RCTs), is good at providing average estimates for treatments or outcomes in the average patient. While EBM has resulted in important clinical guidelines, it does not solve the real clinical quandaries: patients appear for care individually, each may differ in important ways from an RCT cohort, and the physician will wonder each time if following EBM will provide best guidance for this unique patient. This is particularly the case for complex and heterogeneous populations, such as those with congenital heart disease (CHD). Indeed, in congenital cardiology, RCTs are both difficult to conduct and commonly not definitive. The complexity of disease, clinical heterogeneity within lesions, and the small number of patients with specific forms of CHD severely degrade the precision and value of estimates of average treatment effects in the average patient provided by RCTs.

In response to mounting concern about the value of EBM for decision-making, we have previously proposed medicine-based evidence (MBE) as a means of synthesizing all available information and applying it to the individual patient [1]. Briefly, we proposed that whenever a physician needs to decide a patient's treatment plan, a library of patient profiles would be interrogated. A nearest neighbor algorithm would then find "approximate matches," a group of patients who share the greatest similarity with the index case. Some of these matches would and others would not have received a certain treatment or developed a certain outcome, such that specific analyses tailored to the clinical question could be performed within this pool of approximate matches. We envisioned that this approach would represent a major step toward true personalized medicine, as individualization of treatment would shift from today's intrinsically subjective human-driven assessment toward a more objective, data- and model-driven process that is more descriptive, integrative, and predictive.

In their recent article, Li et al [2] introduced CHDmap, an innovative patient similarity network (PSN) designed to prognosticate outcomes among patients with CHD. By leveraging comprehensive clinical and echocardiographic data sets from 4774 surgical cases, the PSN manifests as an interactive, zoomable electronic cartography, wherein each node symbolizes an individual patient, and internode distances delineate their similarity. This user-centric software empowers practitioners to delineate specific patient attributes—such as age, gender, CHD classification, and echocardiographic metrics—tailoring the analysis to the case at hand. The program subsequently uses a k-nearest neighbor algorithm to identify a cohort of closely resembling peers according to the top-k parameter or similarity threshold. This assemblage of approximate matches serves as the foundation for diverse predictive analyses, encompassing variables like hospital length of stay, complications, and survival. This way, CHDmap allows for conducting real-time clinical trials that are specifically tailored to the individual patient, based on historical cases with a similar clinical profile. A key finding from the study by Li et al [2] was the tool's dual capacity: not only did it corroborate clinical intuition in many scenarios, but in specific instances, it prompted a reevaluation of cases, culminating in an enhancement of overall performance across various classification tasks.

The tool has been made publicly available [3] and represents a promising technical rendition of the MBE concept. According to the authors, future generations of the software will be uploaded in time, further expanding the possibilities of CHDmap, including the following: (1) Labeling and visualization of increasingly complex and rare CHD types—currently, only some major subtypes (atrial septal defect, patent foramen ovale, ventricular septal defect, patent ductus arteriosus) are depicted in the map overview; as the underlying data set expands, patients with more complex anatomy may be visualized as well. (2) More granularity in data—in a similar manner, the width of the underlying data set (ie, number of cases) and its depth (ie, number of variables) will likely increase, allowing for more precise matching and examination of more aspects of decision-making. (3) Long-term outcomes—currently, only in-hospital outcomes can be considered within CHDmap, but future generations of the software may allow for long-term outcomes to be analyzed. (4) Comparisons of specific treatment options—once in-depth data on various treatments become available, the optimal treatment for an individual patient may be examined through real-time clinical trials within CHDmap, where outcomes after initiation of various treatments are compared among a group of approximate matches. (5) Artificial intelligence–driven weighting of indicators—the default setting in CHDmap allocates to each indicator the same weighting, whereas physicians can modify these weights based on their prior knowledge; the latter option allows accounting for the fact that weights are likely to differ depending on the clinical setting and the question at hand. With future generations of CHDmap, the authors may implement an artificial intelligence model to dynamically allocate weights to each of the indicators.

CHDmap undeniably signifies a significant stride toward the actualization of MBE. Just as with any statistical methodology, the principles of implementation science will play a pivotal role in optimizing the widespread integration of this tool into clinical practice [4]. Medical practitioners will need to be trained to use these tools correctly and to ensure they are aware of the perks and pitfalls of the PSN (eg, knowing that there is a trade-off between increasing similarity and increasing statistical power or being able to correctly interpret the certainty associated with a specific prediction). Data infrastructure will need to be in place, and continued efforts should be made to establish multicenter clinical registries with in-depth and up-to-date information collection. Furthermore, collaboration between health care professionals and experts in data science will be required to ensure these novel technologies can benefit our patients, taking into account issues regarding data quality and privacy.

Finally, at some point in the future, tools like CHDmap may become routinely used to support team discussions. Rather than replacing the clinician, they should be embraced as assistive technology enhancing overall clinical efficacy. This symbiotic approach serves to harmonize real-life patient care with prior experience and established evidence. This way, we can truly start to achieve the incremental benefits of future generations of physicians learning from previous ones.

Conflicts of Interest

None declared.

References

1. Van den Eynde J, Manlhiot C, Van De Bruaene A, et al. Medicine-based evidence in congenital heart disease: how artificial intelligence can guide treatment decisions for individual patients. Front Cardiovasc Med 2021 Dec;8:798215. [doi: 10.3389/fcvm.2021.798215] [Medline: 34926630]
2. Li H, Zhou M, Sun Y, et al. A patient similarity network (CHDmap) to predict outcomes after congenital heart surgery: development and validation study. JMIR Med Inform 2024 Jan 19;12:e49138. [doi: 10.2196/49138] [Medline: 38297829]
3. CHDmap online interactive tool. Clinical Genetic Test Report Platform. URL: http://chdmap.nbscn.org/ [accessed 2024-04-02]
4. Manlhiot C, Van den Eynde J, Kutty S, Ross HJ. A primer on the present state and future prospects for machine learning and artificial intelligence applications in cardiology. Can J Cardiol 2022 Feb;38(2):169-184. [doi: 10.1016/j.cjca.2021.11.009] [Medline: 34838700]

Abbreviations

CHD: congenital heart disease
EBM: evidence-based medicine
MBE: medicine-based evidence
PSN: patient similarity network
RCT: randomized controlled trial

Edited by C Lovis, S Gardezi; submitted 31.08.23; peer-reviewed by H Li; accepted 10.03.24; published 19.04.24.

Please cite as:
Van den Eynde J
CHDmap: One Step Further Toward Integrating Medicine-Based Evidence Into Practice
JMIR Med Inform 2024;12:e52343
URL: https://medinform.jmir.org/2024/1/e52343
doi:10.2196/52343

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org