Contents

Reviews

Multicriteria Decision-Making in Diabetes Management and Decision Support: Systematic Review (e47701)
Tahmineh Aldaghi, Jan Muzik ................................. 18

Frameworks, Dimensions, Definitions of Aspects, and Assessment Methods for the Appraisal of Quality of Health Data for Secondary Use: Comprehensive Overview of Reviews (e51560)
Jens Declerck, Dipak Kalra, Robert Vander Stichele, Pascal Coorevits .................. 36

Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review (e52967)

Viewpoints

Using a Natural Language Processing Approach to Support Rapid Knowledge Acquisition (e53516)
Taneya Koonce, Dario Giuse, Annette Williams, Mallory Blasingame, Poppy Krump, Jing Su, Nunzia Giuse ................................. 55

The Current Status and Promotional Strategies for Cloud Migration of Hospital Information Systems in China: Strengths, Weaknesses, Opportunities, and Threats Analysis (e52080)
Jian Xu ............................................................................ 61

Implementation Reports

Design and Implementation of an Inpatient Fall Risk Management Information System (e46501)
Ying Wang, Mengyao Jiang, Mei He, Meijie Du ................................................................. 74

A Nationwide Chronic Disease Management Solution via Clinical Decision Support Services: Software Development and Real-Life Implementation Report (e49986)
Mustafa Ulgu, Gokce Laleci Erturkmen, Mustafa Yuksel, Tuncay Namlı, eram Postaci, Mert Gencturk, Yildiray Kabak, A Sinaci, Suat Gonul, Asuman Dogac, Zübeyde Özkan Altunay, Banu Ekinci, Sahin Aydin, Suayip Birinci .................................................. 82
Ten Years of Experience With a Telemedicine Platform Dedicated to Health Care Personnel: Implementation Report (e42847)
Claudio Azzolini, Elias Premi, Simone Donati, Andrea Falco, Aldo Torreggiani, Francesco Sicurello, Andreina Baj, Lorenzo Azzi, Alessandro Orro, Giovanni Porta, Giovanna Azzolini, Marco Sorrentino, Paolo Melillo, Francesco Testa, Francesca Simonelli, Gianfranco Giardina, Umberto Paolucci. 97

Learnings From Implementation of Technology-Enabled Mental Health Interventions in India: Implementation Report (e47504)
Sudha Kallakuri, Sridevi Gara, Mahesh Godi, Sandhya Yatirajula, Srilatha Paslawar, Mercian Daniel, David Peiris, Pallab Maulik. 111

Original Papers

Value of Electronic Health Records Measured Using Financial and Clinical Outcomes: Quantitative Study (e52524)
Shikha Modi, Sue Feldman, Eta Berner, Benjamin Schooley, Allen Johnston. 126

Additional Value From Free-Text Diagnoses in Electronic Health Records: Hybrid Dictionary and Machine Learning Classification Study (e49007)
Tarun Mehra, Tobias Wekhof, Dagmar Keller. 139

BERT-Based Neural Network for Inpatient Fall Detection From Electronic Medical Records: Retrospective Cohort Study (e48995)
Cheligeer Cheligeer, Guosong Wu, Seungwon Lee, Jie Pan, Danielle Southern, Elliot Martin, Natalie Sapio, Cathy Eastwood, Hude Quan, Yuan Xu. 156

Prediction of Antibiotic Resistance in Patients With a Urinary Tract Infection: Algorithm Development and Validation (e51326)
Nevruz Ihanli, Se Park, Jaewoong Kim, Jee Ryu, Ahmet Yardimci, Dukyong Yoon. 183

The Implementation of an Electronic Medical Record in a German Hospital and the Change in Completeness of Documentation: Longitudinal Document Analysis (e47761)
Florian Wurster, Marina Beckmann, Natalia Ceco-Stabel, Kerstin Dittmer, Till Hansen, Julia Jaschke, Juliane Köberlein-Neu, Mi-Ran Okum, Carsten Rusniok, Holger Pfaff, Ute Karbach. 194

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 (e53002)
Siyi Zhan, Liping Ding, Hui Li, Aonan Su. 205

Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation Use Case: Implementation and Evaluation Study (e47744)
Boshu Ru, Arthur Sillah, Kaushal Desai, Sheenu Chandwani, Lixia Yao, Smita Kothari. 219

Enhancing Health Equity by Predicting Missed Appointments in Health Care: Machine Learning Study (e48273)
Yi Yang, Sameh Madanani, David Parry. 247

Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study (e51925)
Anne de Hond, Marieke van Buchem, Claudio Fanconi, Mohana Roy, Douglas Blayney, Ilse Kant, Ewout Steyerberg, Tina Hernandez-Boussard. 264
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 (e42271)

Angie Li, Sarah Mullin, Peter Elkin
Predicting Hypoxia Using Machine Learning: Systematic Review

Lena Pigat¹, MPH; Benjamin P Geisler¹, MD, MPH; Seyedmostafa Sheikhalishahi¹, PhD; Julia Sander¹, PhD; Mathias Kaspar¹, PhD; Maximilian Schmutz¹,², MD; Sven Olaf Rohr¹, MD; Carl Mathis Wild¹,³, MD; Sebastian Goss¹, MD; Sarra Zaghdoudi¹, MSc; Ludwig Christian Hinske¹,⁴, Prof Dr

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 hypoxic 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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample size n</th>
<th>Clinical variables, n</th>
<th>Prediction end point</th>
<th>Model</th>
<th>Performance</th>
<th>External validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annapragada et al [15] (2021)</td>
<td>2435</td>
<td>1</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt; &lt;8% within the next 5 and 30 min (occurrence and magnitude of hypoxicem events)</td>
<td>LSTM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>• PPV&lt;sup&gt;c&lt;/sup&gt;: 0.94 • Sensitivity: 0.80 • Specificity: 0.99</td>
<td>Yes</td>
</tr>
<tr>
<td>Chen et al [16] (2021)</td>
<td>57,171</td>
<td>21</td>
<td>SaO&lt;sub&gt;2&lt;/sub&gt; &lt;93% within the next 5 min</td>
<td>GBT&lt;sup&gt;a&lt;/sup&gt;</td>
<td>• AUROC&lt;sup&gt;e&lt;/sup&gt;: 0.89</td>
<td>Yes</td>
</tr>
<tr>
<td>ElMoaqet et al [17] (2014)</td>
<td>119</td>
<td>1</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt; ≤89% within the next 20 and 60 s</td>
<td>Lin&lt;sup&gt;g&lt;/sup&gt;</td>
<td>• AUROC: 0.93</td>
<td>No</td>
</tr>
<tr>
<td>Erion et al [18] (2017)</td>
<td>57,173</td>
<td>1</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt; ≤92% within the next 5 min</td>
<td>LSTM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>• LSTM AUROC: 0.87 • GBT AUROC: 0.86</td>
<td>No</td>
</tr>
<tr>
<td>Geng et al [23] (2018)</td>
<td>308</td>
<td>3</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt; &lt;90% for any duration during the endoscopic procedure</td>
<td>LR&lt;sup&gt;b&lt;/sup&gt;</td>
<td>• AUROC: 0.76</td>
<td>No</td>
</tr>
<tr>
<td>Geng et al [24] (2019)</td>
<td>220</td>
<td>3</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt; &lt;90% for any duration during the endoscopy procedure</td>
<td>ANN&lt;sup&gt;i&lt;/sup&gt;</td>
<td>• AUROC: 0.80</td>
<td>No</td>
</tr>
<tr>
<td>Lam et al [19] (2022)</td>
<td>39,630</td>
<td>26</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt; &lt;91% and &lt;96% after algorithm evaluation and any time during hospitalization</td>
<td>XGB&lt;sup&gt;b&lt;/sup&gt;</td>
<td>• XGB AUROC: 0.64 • RNN AUROC: 0.64</td>
<td>Yes</td>
</tr>
<tr>
<td>Lundberg et al [20] (2018)</td>
<td>36,232</td>
<td>&gt;65</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt; ≤92% initial status and within the next 5 min</td>
<td>GBM&lt;sup&gt;i&lt;/sup&gt;</td>
<td>• AUROC: 0.90</td>
<td>No</td>
</tr>
<tr>
<td>Ren et al [21] (2022)</td>
<td>17,818</td>
<td>3</td>
<td>PaO&lt;sub&gt;2&lt;/sub&gt;/FiO&lt;sub&gt;2&lt;/sub&gt; ≤150 at any time during ventilation</td>
<td>NN&lt;sup&gt;o&lt;/sup&gt;</td>
<td>• NN AUROC: 0.83 • LR AUROC: 0.81</td>
<td>Yes</td>
</tr>
<tr>
<td>Sippl et al [25] (2017)</td>
<td>620</td>
<td>17, RF&lt;sup&gt;f&lt;/sup&gt; and NN used subsets of 6 and 7</td>
<td>Presence and severity of temporary oxygen desaturation during anesthesia induction and intubation based on expert annotations</td>
<td>NN&lt;sup&gt;f&lt;/sup&gt;</td>
<td>• NN sensitivity: 0.74 • NN specificity: 0.93 • RF sensitivity: 0.35 • RF specificity: 0.99</td>
<td>No</td>
</tr>
<tr>
<td>Statsenko et al [26] (2022)</td>
<td>605</td>
<td>2D and 3D diagnostic images of the chest</td>
<td>Markers of systemic oxygenation: functional (HR&lt;sup&gt;3&lt;/sup&gt;, BR&lt;sup&gt;3&lt;/sup&gt;, SBP&lt;sup&gt;3&lt;/sup&gt;, and DBP&lt;sup&gt;3&lt;/sup&gt;) and biochemical findings (SpO&lt;sub&gt;2&lt;/sub&gt;, serum potassium level, and AG&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>CNN&lt;sup&gt;v&lt;/sup&gt;</td>
<td>• MAE&lt;sup&gt;w&lt;/sup&gt;: mean 7.941% (SD 4.131%)</td>
<td>No</td>
</tr>
<tr>
<td>Reference</td>
<td>Sample size n</td>
<td>Clinical variables, n</td>
<td>Prediction end point</td>
<td>Model</td>
<td>Performance</td>
<td>External validation</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------</td>
<td>-----------------------</td>
<td>---------------------------------------------------------------</td>
<td>------------</td>
<td>-------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Xia et al [22] (2022)</td>
<td>14,777</td>
<td>29</td>
<td>PaO₂ &lt;60 mm Hg after extubating</td>
<td>• RF</td>
<td>• AUROC: 0.792</td>
<td>No</td>
</tr>
</tbody>
</table>

*a*SpO₂: peripheral oxygen saturation.  
*b*LSTM: long short-term memory.  
*c*PPV: positive predictive value.  
*d*SaO₂: arterial oxygen saturation.  
*e*GBT: gradient boosted tree.  
*f*AUROC: area under the receiver operating characteristics.  
*g*Lin: linear regression.  
*h*L: logistic regression.  
*i*ANN: artificial neural network.  
*j*XGB: extreme gradient boosting.  
*k*RNN: recurrent neural network.  
*l*GBM: gradient boosting machine.  
*m*PaO₂: partial pressure of oxygen.  
*n*FiO₂: 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=5) [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 (SpO₂) 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 SpO₂ 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\textsubscript{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\textsubscript{2} data applied a long short-term memory (LSTM) algorithm, 1 of which was able to predict the detailed trend of the SpO\textsubscript{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].
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 (e.g., 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 SpO₂ 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

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
Review

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

Tahmineh Aldaghi1*, MSc; Jan Muzik2*, PhD

1Spin-off Companies and Research Results Commercialization Center, First Faculty of Medicine, Charles University, Prague, Czech Republic
2Department 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
Studnická 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 etiology 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.
Methods

Search Strategies
A query was carried out on PubMed, Elsevier, Embase, MEDLINE, Scopus, MBC, Springer, IEE, 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].

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 Réalité (ELECTRE).
(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 Réalité; 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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Methods</th>
<th>Objective</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maruthur et al [14]</td>
<td>AHP</td>
<td>Select oral T2D medications</td>
<td>Sitagliptin, sulfonylureas, and pioglitazone</td>
</tr>
<tr>
<td>Eghbali-Zarch et al [29]</td>
<td>SWARA(^c) method, ratio analysis, and the FMULTIMOORA(^d) method</td>
<td>Choose the pharmacological treatment for T2D</td>
<td>Metformin should be used as the first-line medication, followed by sulfonylurea, glucagon-like peptide-1 receptor agonist, dipeptidyl peptidase-4 inhibitor, and insulin</td>
</tr>
<tr>
<td>Eghbali-Zarch et al [28]</td>
<td>WASPAS(^e), entropy, and CRITIC(^f)</td>
<td>Determine the final ranking of the medications</td>
<td>Proposed a model to help endocrinologist to choose the best medicine</td>
</tr>
<tr>
<td>Zhang et al [30]</td>
<td>TOPSIS(^g)</td>
<td>Ranking of diabetes medicines</td>
<td>CDSS(^h) can assist young doctors and nonspecialty physicians with medication prescriptions</td>
</tr>
<tr>
<td>Maruthur et al [31]</td>
<td>AHP</td>
<td>Select oral T2D medications</td>
<td>AHP will aid, support, and enhance the ability of decision makers to make evidence-based informed decisions consistent with their values and preferences</td>
</tr>
<tr>
<td>Nag and Helal [32]</td>
<td>Fuzzy AHP and AHP</td>
<td>Classification of diabetic medications</td>
<td>Fuzzy AHP model can better handle the ambiguity of decision makers</td>
</tr>
<tr>
<td>Chen et al [33]</td>
<td>Entropy</td>
<td>Choose pharmaceuticals</td>
<td>AGI(^i), DPP4(^i), MET(^k), Glinide, SU(^l), and TZD(^m)</td>
</tr>
<tr>
<td>Wang et al [34]</td>
<td>AHP and ANP(^p)</td>
<td>Combine different clinical, economic, and medical decision-making elements</td>
<td>Modifying one’s lifestyle, taking metformin, and receiving insulin injections</td>
</tr>
<tr>
<td>Bao et al [35]</td>
<td>MCDA(^o)</td>
<td>Assess medicine for diabetes</td>
<td>Five DPP4 inhibitors was valuable</td>
</tr>
<tr>
<td>Onar and Ibil [36]</td>
<td>Fuzzy AHP</td>
<td>Considered the best oral antidiabetic</td>
<td>Proposed a decision support system</td>
</tr>
<tr>
<td>Zhang et al [37]</td>
<td>MCDA</td>
<td>Examine the Mudan Granules</td>
<td>The new medication was acceptable</td>
</tr>
<tr>
<td>Cai et al [38]</td>
<td>AHP</td>
<td>Evaluate strains of the efficacy of the LAB(^p) with possible antidiabetic capabilities</td>
<td>Potential antidiabetic effect</td>
</tr>
<tr>
<td>Sekar et al [39]</td>
<td>Fuzzy PROMETHEE(^q)</td>
<td>Choose the best course of therapy</td>
<td>Giving the high peace of treatment to the most affected people</td>
</tr>
<tr>
<td>Mühlbacher et al [40]</td>
<td>AHP and BWS(^r)</td>
<td>Evaluate patients’ preferences for various T2D treatment parameters</td>
<td>Proposed a model</td>
</tr>
<tr>
<td>Mahat and Ahmad [41]</td>
<td>Fuzzy AHP</td>
<td>Identify and choose the most efficient thermal massage treatment session</td>
<td>Number of therapy sessions (per day) was the most important factor</td>
</tr>
<tr>
<td>Pan et al [42]</td>
<td>Fuzzy AHP</td>
<td>Determine the weights of the various physiological factors</td>
<td>The mathematical model of exercise rehabilitation program for patients with diabetes was established</td>
</tr>
<tr>
<td>Rani et al [43]</td>
<td>COPRAS(^s)</td>
<td>Select T2D medication treatment</td>
<td>Developed a new formula-based PFS(^s) and evaluated its feasibility by applying the model on selecting the T2D pharmacological therapy</td>
</tr>
<tr>
<td>Balubaid and Basheikh [44]</td>
<td>AHP</td>
<td>Developed a mathematical decision-making model that prioritizes the available diabetes medication based on criteria</td>
<td>Metformin, pioglitazone, sitagliptin, and glimepiride were ranked first, second, third, and fourth, respectively</td>
</tr>
<tr>
<td>Mühlbacher et al [45]</td>
<td>AHP and BWS</td>
<td>Examine the key patient-related decision criteria involved in the medicinal treatment of T2D</td>
<td>For oral antidiabetes-treated patient groups and insulin-treated patient groups, HbA1(^c) level, delay of insulin therapy, and occurrence of hypoglycemia were ranked first, second, and third, respectively</td>
</tr>
</tbody>
</table>

\(^a\)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.
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.

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

<sup>a</sup>TOPSIS: technique for order of preference by similarity to ideal solution.  
<sup>b</sup>BP: blood pressure.  
<sup>c</sup>ACR: albumin creatinine ratio.  
<sup>d</sup>AHP: analytic hierarchy process.  
<sup>e</sup>FAHP: fuzzy analytic hierarchy process.  
<sup>f</sup>GDM: gestational diabetes mellitus.  
<sup>g</sup>N/A: not applicable.  
<sup>h</sup>T2D: type 2 diabetes.  
<sup>i</sup>FBS: fasting blood sugar.  
<sup>j</sup>PRF: physical risk factors.  
<sup>k</sup>DIBAR: Created Diabetes Risk Assessment.  
<sup>l</sup>FRBS: fuzzy rule-based systems.  
<sup>m</sup>EDAS: evaluation based on distance for average solution.  
<sup>n</sup>MCDM: multicriteria decision-making.  
<sup>o</sup>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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Methods</th>
<th>Objective</th>
<th>Criteria</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaikwad et al [56]</td>
<td>AHP</td>
<td>Recommend a particular ice cream for patients with diabetes</td>
<td>Sugar, cholesterol, dietary fiber, and proteins</td>
<td>Ben &amp; Jerry’s Butter Pecan was enriched with all 4 criteria</td>
</tr>
<tr>
<td>Sharawat and Dubey [57]</td>
<td>AHP</td>
<td>Find out the best diet for a patient with diabetes among 3 alternatives: solid food, liquid food, and fluid food</td>
<td>Calories, body fat, healthy carbs, and dietary needs</td>
<td>Solid food was selected as the best</td>
</tr>
<tr>
<td>Santosso et al [58]</td>
<td>Fuzzy AHP</td>
<td>Designed a new yogurt product for patients with diabetes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Zadeh et al [59]</td>
<td>AHP</td>
<td>Proposed a personalized meal-planning strategy</td>
<td>N/A</td>
<td>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</td>
</tr>
<tr>
<td>Gulint and Kadam [60]</td>
<td>AHP and TOPSIS</td>
<td>Recommended shakes and ice cream for patients with diabetes</td>
<td>Sugar, cholesterol, carbs, fat, protein, and dietary fiber</td>
<td>Selected a type of ice cream that satisfies all criteria</td>
</tr>
<tr>
<td>Gaikwad et al [61]</td>
<td>ANP</td>
<td>Recommendation of a particular ice cream</td>
<td>Sugar, calories, cholesterol, and proteins</td>
<td>Selected a type of ice cream that satisfies all criteria</td>
</tr>
<tr>
<td>Gaikwad et al [62]</td>
<td>AHP</td>
<td>Recommendation of a particular ice cream</td>
<td>N/A</td>
<td>Proposed a model combination of AHP-GA and AHP-CI to recommend an ice cream to patients with diabetes</td>
</tr>
<tr>
<td>Gaikwad et al [63]</td>
<td>AHP</td>
<td>Recommendation of a particular ice cream</td>
<td>Sugar, protein, cholesterol, and dietary fiber</td>
<td>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</td>
</tr>
</tbody>
</table>

aAHP: analytic hierarchy process.
bN/A: not applicable.
cTOPSIS: technique for order of preference by similarity to ideal solution.
dANP: analytical network process.
AHP-CI: analytic hierarchy process–cohort intelligence.

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.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Method</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gupta et al [64]</td>
<td>TOPSIS^a, VIKOR^b, PROMETHEE II^c</td>
<td>Assess current mHealth^d applications for T2D^e, including Glucose Buddy, mySugr, Diabetes: M, Blood Glucose Tracker, and OneTouch Reveal</td>
</tr>
<tr>
<td>Wang et al [65]</td>
<td>ANP^f and CRITIC^g</td>
<td>Assess the influence of social support on T2DM^h self-management</td>
</tr>
<tr>
<td>Mishra et al [66]</td>
<td>AHP^i</td>
<td>Created and used the SCP^j assessment methodology for Indian diabetes clinic</td>
</tr>
<tr>
<td>Mishra [67]</td>
<td>AHP</td>
<td>Developed a customized service quality assessment model for diabetes care</td>
</tr>
<tr>
<td>Mishra [68]</td>
<td>Fuzzy TOPSIS</td>
<td>Proposed 3 alternatives for the placement of a diabetes clinic using the SLP^k method</td>
</tr>
<tr>
<td>Byun et al [69]</td>
<td>AHP</td>
<td>Improving the treatment compliance of patients with diabetes</td>
</tr>
<tr>
<td>Mehrotra and Kim [70]</td>
<td>New multicriteria, robust weighted-sum methodology</td>
<td>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</td>
</tr>
<tr>
<td>Haji et al [71]</td>
<td>AHP and TOPSIS</td>
<td>Create a model that can prioritize and pick the optimal criterion for optimizing insulin safety</td>
</tr>
<tr>
<td>Suka et al [72]</td>
<td>AHP</td>
<td>Described a clinical decision support system that enhance dynamic decision-making</td>
</tr>
<tr>
<td>Fico et al [73]</td>
<td>AHP</td>
<td>Selected the best tool for screening and managing T2D</td>
</tr>
<tr>
<td>Long and Centor [74]</td>
<td>AHP</td>
<td>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</td>
</tr>
<tr>
<td>Gajdoš et al [75]</td>
<td>TOPSIS</td>
<td>Proposed a concept of chronic care management, which could increase effectiveness and reduce the cost of health care provided to patients with T2D</td>
</tr>
<tr>
<td>Gupta et al [76]</td>
<td>CODAS-FAHP^n and MOORA-FAHP^o</td>
<td>Assess the usability of mHealth applications to monitor T2D by developing 2 hybrid decision-making methods</td>
</tr>
<tr>
<td>Chang et al [77]</td>
<td>Delphi-AHP</td>
<td>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</td>
</tr>
</tbody>
</table>

^aTOPSIS: technique for order of preference by similarity to ideal solution.  
^bVIKOR: ViseKriterijumska Optimizacija I Kompromisno Resenje.  
^cPROMETHEE II: preference ranking organization method for enrichment of evaluation II.  
^dT2D: type 2 diabetes.  
^eT2D: type 2 diabetes mellitus.  
^fANP: analytical network process.  
^gCRITIC: Criteria Importance Through Intercriteria Correlation  
^hT2DM: type 2 diabetes mellitus.  
^iAHP: analytic hierarchy process.  
^jSCP: Supply Chain Partnership.  
^kSLP: Systematic Layout Planning.  
^lHbA1c: hemoglobin A1c.  
^mLDL: low-density lipoprotein.  
^nCODAS-FAHP: combine distance-based assessment-fuzzy AHP.  
^oMOORA-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.

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

<sup>a</sup>TOPSIS: technique for order of preference by similarity to ideal solution.  
<sup>b</sup>BP: blood pressure.  
<sup>c</sup>MCDM: multicriteria decision-making.  
<sup>d</sup>GFR: estimated glomerular filtration rate.  
<sup>e</sup>CKD: chronic kidney disease.  
<sup>f</sup>DR: diabetic retinopathy.  
<sup>g</sup>AUC: area under the curve.  
<sup>h</sup>TPR: true positive rate.  
<sup>i</sup>FNR: false negative rate.  
<sup>j</sup>FPR: false positive rate.  
<sup>k</sup>TNR: true negative rate.  
<sup>l</sup>VIKOR: ViseKriterijumska Optimizacija I Kompromisno Resenje.  
<sup>m</sup>HDL: high-density lipoprotein.  
<sup>n</sup>AHP: analytic hierarchy process.  
<sup>o</sup>DBP: diastolic blood pressure.  
<sup>p</sup>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 smooth 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 - medinfor_m_v12i1e47701_app1.docx ]

References


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 Réalité
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
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 Declerck1,2, MSc; Dipak Kalra1,2, Prof Dr; Robert Vander Stichele3, Prof Dr; Pascal Coorevits1, Prof Dr
1Department of Public Health and Primary Care, Unit of Medical Informatics and Statistics, Ghent University, Ghent, Belgium
2The European Institute for Innovation through Health Data, Ghent, Belgium
3Faculty 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.


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.

https://medinform.jmir.org/2024/1/e51560 JMIR Med Inform 2024 | vol. 12 | e51560 | p.37 (page number not for citation purposes)
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].

<table>
<thead>
<tr>
<th>Data quality dimension and definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Completeness: the extent to which data are present</td>
</tr>
<tr>
<td>• Consistency: the extent to which data satisfy constraints</td>
</tr>
<tr>
<td>• Correctness: the extent to which data are true and unbiased</td>
</tr>
<tr>
<td>• Timeliness: the extent to which data are promptly processed and up to date</td>
</tr>
<tr>
<td>• Stability: the extent to which data are comparable among sources and over time</td>
</tr>
<tr>
<td>• Contextualization: the extent to which data are annotated with acquisition context</td>
</tr>
<tr>
<td>• Representativeness: the extent to which data are representative of intended use</td>
</tr>
<tr>
<td>• Trustworthiness: the extent to which data can be trusted based on the owner’s reputation</td>
</tr>
<tr>
<td>• Uniqueness: the extent to which data are not duplicated</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completeness</strong></td>
<td>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] defined completeness as the presence of...</td>
</tr>
</tbody>
</table>

- **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. 

- **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. 

- **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. 

- **Completeness**  
  **EHR:** electronic health record.
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 codings 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>
<thead>
<tr>
<th>Assessment M&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Assessment technique in reviews</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Linkages—other data sets</td>
<td>- Percentage of eligible population included in the data set.</td>
</tr>
<tr>
<td>M2</td>
<td>Comparison of distributions</td>
<td>- Difference in means and other statistics.</td>
</tr>
<tr>
<td>M3</td>
<td>Case duplication</td>
<td>- Number and percentage or cases with &gt;1 record.</td>
</tr>
<tr>
<td>M4</td>
<td>Completeness of variables</td>
<td>- Percentage of cases with complete observations of each variable.</td>
</tr>
<tr>
<td>M5</td>
<td>Completeness of cases</td>
<td>- Percentage of cases with complete observations for all variables.</td>
</tr>
<tr>
<td>M6</td>
<td>Distribution comparison</td>
<td>- Distributions or summary statistics of aggregated data from the data set are compared with the expected distributions for the clinical concepts of interest.</td>
</tr>
<tr>
<td>M7</td>
<td>Gold standard</td>
<td>- A data set drawn from another source or multiple sources is used as a gold standard.</td>
</tr>
</tbody>
</table>
| 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<sup>b</sup>         | M:I                           | - 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. |
<table>
<thead>
<tr>
<th>Assessment M&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Assessment technique in reviews</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>M24</td>
<td>Not specified</td>
<td>• (I=\text{ratio of violations of specific consistency type to the total number of consistency checks.})</td>
</tr>
<tr>
<td>M25</td>
<td>Validity check</td>
<td>• Data in the data set are assessed using various techniques that determine of the values &quot;make sense.&quot;</td>
</tr>
</tbody>
</table>
| 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.                                                                                                          |

<sup>a</sup>M: method.
<sup>b</sup>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.

Multimedia Appendix 2
Data sources, data quality aspects, and definitions reported in the 22 publications included in the review.

Multimedia Appendix 3
The frequency of all dimensions with definitions in each review and assessment methods per dimension.

Multimedia Appendix 4
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.

References


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:
Viewpoint

Using a Natural Language Processing Approach to Support Rapid Knowledge Acquisition

Taneya Y Koonce¹, MSLS, MPH; Dario A Giuse², MS, Dr Ing; Annette M Williams¹, MLS; Mallory N Blasingame¹, MA, MSIS; Poppy A Krump¹, MSIS; Jing Su¹, MS, MSIS, MD; Nunzia B Giuse¹,², MLS, MD

¹Center for Knowledge Management, Vanderbilt University Medical Center, Nashville, TN, United States
²Department 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:

\[
\text{Ratio} = \frac{\frac{n_1}{T}}{\frac{n_2}{T}}
\]

where \( T \) = total population size, \( n_1 = \) number of patients with both concept 1 and concept 2, \( n_2 = \) 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.

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

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

<table>
<thead>
<tr>
<th>CUI</th>
<th>Name</th>
<th>Expect</th>
<th>Actual</th>
<th>Ratio</th>
<th>Pop. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4076686</td>
<td>supine hypertension</td>
<td>49,565</td>
<td>2,721</td>
<td>54.898</td>
<td>2,916</td>
</tr>
<tr>
<td>C0393911</td>
<td>autonomic failure</td>
<td>50,006</td>
<td>2,698</td>
<td>53.953</td>
<td>2,942</td>
</tr>
<tr>
<td>C1868528</td>
<td>autonomic hypotension</td>
<td>2,363</td>
<td>125</td>
<td>52.907</td>
<td>139</td>
</tr>
<tr>
<td>C2315246</td>
<td>pandysautonomia</td>
<td>0.714</td>
<td>36</td>
<td>50.428</td>
<td>42</td>
</tr>
<tr>
<td>C0877382</td>
<td>delayed renal graft function</td>
<td>0.544</td>
<td>26</td>
<td>47.801</td>
<td>32</td>
</tr>
<tr>
<td>C3662086</td>
<td>multisystem degeneration of autonomic nervous system</td>
<td>0.629</td>
<td>30</td>
<td>47.702</td>
<td>37</td>
</tr>
<tr>
<td>C0037019</td>
<td>shy-drager syndrome</td>
<td>8,686</td>
<td>410</td>
<td>52.764</td>
<td>511</td>
</tr>
<tr>
<td>C1959798</td>
<td>baroreflex failure syndrome</td>
<td>5,337</td>
<td>237</td>
<td>44.405</td>
<td>314</td>
</tr>
<tr>
<td>C0393571</td>
<td>multisystem atrophy</td>
<td>14,482</td>
<td>631</td>
<td>43.572</td>
<td>852</td>
</tr>
<tr>
<td>C0877157</td>
<td>pancreas transplant rejection</td>
<td>1,377</td>
<td>56</td>
<td>40.674</td>
<td>81</td>
</tr>
<tr>
<td>C5395071</td>
<td>autoimmune gaugionopathy</td>
<td>0.153</td>
<td>6</td>
<td>39.222</td>
<td>9</td>
</tr>
<tr>
<td>C1963757</td>
<td>dopamine dysregulation syndrome</td>
<td>0.561</td>
<td>21</td>
<td>37.439</td>
<td>33</td>
</tr>
<tr>
<td>C0920233</td>
<td>pisa syndrome</td>
<td>1.122</td>
<td>40</td>
<td>35.656</td>
<td>66</td>
</tr>
</tbody>
</table>

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 “multisystem 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, searches 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
22. PhenoCEx (Extended). PhenoWAS Resources. URL: https://phewascatalog.org/phenoCEx [accessed 2023-12-13]
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
PhenoWAS: phenome-wide association study
UMLS: Unified Medical Language System
VUMC: Vanderbilt University Medical Center

©Taneya Y Koonce, Dario A Giuse, Annette M Williams, Mallory N Blasingame, Poppy A Krump, Jing Su, Nunzia B Giuse. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Abstract

Background: In the 21st century, Chinese hospitals have witnessed innovative medical business models, such as online diagnosis and treatment, cross-regional multidisciplinary 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
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: (TT (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) were generated for literature review and Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis.
Textbox 1. Inclusion and exclusion criteria for literature review.

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article type: fully retrievable</td>
<td>Article type: nonretrievable</td>
</tr>
<tr>
<td>Language: English, Chinese</td>
<td>Language: other languages</td>
</tr>
<tr>
<td>Nationality of the first author: Chinese (including Hong Kong and Taiwan)</td>
<td>Nationality of the first author: other countries</td>
</tr>
<tr>
<td>Article topic: the research, development, and application of cloud technology in Chinese hospitals</td>
<td>Article topic: other topics</td>
</tr>
<tr>
<td>Publication date: from January 1, 2007, to February 21, 2023</td>
<td>Publication date: before January 1, 2007; after February 21, 2023</td>
</tr>
</tbody>
</table>

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.
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.

<table>
<thead>
<tr>
<th>Research methods</th>
<th>Hospitals, n (%)</th>
<th>Universities or colleges, n (%)</th>
<th>Associations or companies, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>System research and development</td>
<td>13 (14)</td>
<td>14 (15)</td>
<td>N/A a</td>
</tr>
<tr>
<td>Cloud platform construction</td>
<td>5 (5)</td>
<td>5 (5)</td>
<td>N/A</td>
</tr>
<tr>
<td>Technical research</td>
<td>2 (2)</td>
<td>8 (9)</td>
<td>N/A</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical research</td>
<td>14 (15)</td>
<td>7 (7)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Case control study</td>
<td>11 (12)</td>
<td>2 (2)</td>
<td>N/A</td>
</tr>
<tr>
<td>Summary of experience</td>
<td>3 (3)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Cohort study</td>
<td>2 (2)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Literature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retrospective study</td>
<td>4 (4)</td>
<td>2 (2)</td>
<td>N/A</td>
</tr>
<tr>
<td>Standard study</td>
<td>N/A</td>
<td>1 (1)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

aN/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.
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% 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 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

**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.

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 (e.g., 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


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]


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/2022/2-10/2/MMDE0MDVIMTY5NjM2Mw.html [accessed 2023-10-03]
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
Design and Implementation of an Inpatient Fall Risk Management Information System

Wang et al

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.
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-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-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 $\times$ 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 $\times$ 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 $\times$ the total number of environment safety check items.

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.**

<table>
<thead>
<tr>
<th>Category</th>
<th>Before implementation (total bed days=1,303,094)</th>
<th>After implementation (total bed days=1,323,667)</th>
<th>Chi-square (df)</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-risk patients’ reports</td>
<td>1036 (0.8)</td>
<td>3007 (2.3)</td>
<td>931.7 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fall incident reports</td>
<td>23 (0.02)</td>
<td>11 (0.01)</td>
<td>4.4 (1)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
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.

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. 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.

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.

Table 2. Results of fall-related injuries.

<table>
<thead>
<tr>
<th>Cases of fall-related injury, n</th>
<th>No injury</th>
<th>Minor</th>
<th>Moderate</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before implementation</td>
<td>15</td>
<td>28</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>After implementation</td>
<td>20</td>
<td>13</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3. Comparison of fall-related process indicators.

<table>
<thead>
<tr>
<th>Process</th>
<th>Before implementation (n=1250), n (%)</th>
<th>After implementation (n=1806), n (%)</th>
<th>Chi-square (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall risk assessment</td>
<td>1056 (84.48)</td>
<td>1709 (95.73)</td>
<td>88 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fall-related health education</td>
<td>1107 (88.56)</td>
<td>1769 (97.95)</td>
<td>117.5 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fall-related shift handoff</td>
<td>1114 (89.12)</td>
<td>1767 (97.84)</td>
<td>104 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Environment safety</td>
<td>1127 (90.16)</td>
<td>1796 (99.45)</td>
<td>153 (1)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
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


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
A Nationwide Chronic Disease Management Solution via Clinical Decision Support Services: Software Development and Real-Life Implementation Report

Mustafa Mahir Ulgu1, MD, PhD; Gokce Banu Laleci Erturkmen2, PhD; Mustafa Yuksel2, PhD; Tuncay Namli2, PhD; Şenan Postacı2, MSc; Mert Gencturk2, PhD; Yildiray Kabak2, PhD; A Anil Sinaci2, PhD; Suat Gonul2, PhD; Asuman Dogac2, PhD; Zübeyde Özkan Altunay1, MD; Banu Ekinci1, MD; Sahin Aydin1, MSc; Suayip Birinci1, MD

1Ministry of Health Turkey, Ankara, Turkey
2Software 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 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.
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, 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 $\text{A}_1^c$ ($\text{HbA}_1^c$), 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:
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, HbA1c, and blood pressure.

The overall system architecture of the DMP is depicted in Figure 2.
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.

https://medinform.jmir.org/2024/1/e49986

(JMIR Med Inform 2024 | vol. 12 | e49986 | p.86)
• 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 m².

• 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.
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 Türiksat 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.
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 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.

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.

<table>
<thead>
<tr>
<th>Module</th>
<th>Screening (n=45,166,536), n (%)</th>
<th>Monitoring (n=28,548,733), n (%)</th>
<th>Total (n=73,715,269), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension</td>
<td>13,857,594 (30.7)</td>
<td>12,046,449 (42.2)</td>
<td>25,904,043 (35.1)</td>
</tr>
<tr>
<td>Obesity</td>
<td>18,029,994 (39.9)</td>
<td>800,480 (2.8)</td>
<td>18,830,474 (25.5)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>8,914,193 (19.7)</td>
<td>5,071,646 (17.8)²</td>
<td>13,985,839 (19.0)</td>
</tr>
<tr>
<td>CVD³ risk</td>
<td>4,364,755 (9.7)</td>
<td>9,182,814 (32.2)</td>
<td>13,547,569 (18.4)</td>
</tr>
<tr>
<td>Older adult</td>
<td>N/A²</td>
<td>1,447,344 (5.1)</td>
<td>1,447,344 (2.0)</td>
</tr>
</tbody>
</table>

¹Only the patients monitored in primary care are listed; advanced obesity cases (a BMI over 40 kg/m² or a BMI between 30 and 40 kg/m² supported with additional comorbidities) are monitored in secondary and tertiary care.

²CVD: cardiovascular disease.

³N/A: not applicable.
### Table 2. Coverage rate of citizens in target population lists.

<table>
<thead>
<tr>
<th>Module and encounter type</th>
<th>All citizens in target population, n</th>
<th>Screened and monitored patients, n</th>
<th>Coverage rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hypertension</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td>48,443,467</td>
<td>10,820,774</td>
<td>22.3</td>
</tr>
<tr>
<td>Monitoring</td>
<td>14,943,378</td>
<td>4,083,057</td>
<td>27.3</td>
</tr>
<tr>
<td><strong>Obesity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td>59,956,288</td>
<td>14,640,013</td>
<td>24.4</td>
</tr>
<tr>
<td>Monitoring</td>
<td>769,654(^a)</td>
<td>383,920</td>
<td>49.9</td>
</tr>
<tr>
<td><strong>Diabetes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td>27,450,172</td>
<td>6,486,947</td>
<td>23.6</td>
</tr>
<tr>
<td>Monitoring</td>
<td>7,588,543</td>
<td>2,472,585</td>
<td>32.6</td>
</tr>
<tr>
<td><strong>CVD(^b) risk</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td>17,276,617</td>
<td>3,319,070</td>
<td>19.2</td>
</tr>
<tr>
<td>Monitoring</td>
<td>17,759,500</td>
<td>5,078,665</td>
<td>28.6</td>
</tr>
<tr>
<td><strong>Older adult</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monitoring</td>
<td>8,770,474</td>
<td>1,056,766</td>
<td>12.0</td>
</tr>
</tbody>
</table>

\(^a\)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.

<table>
<thead>
<tr>
<th>City</th>
<th>Total populationa</th>
<th>Rankb</th>
<th>Number of encounters</th>
<th>Number of patients</th>
<th>Average age of patients (years)</th>
<th>Average duration (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Istanbul</td>
<td>15,907,951</td>
<td>1</td>
<td>4,974,972</td>
<td>1,286,640</td>
<td>50.5</td>
<td>1.21</td>
</tr>
<tr>
<td>Ankara</td>
<td>5,782,285</td>
<td>2</td>
<td>4,918,745</td>
<td>1,084,797</td>
<td>51.7</td>
<td>1.13</td>
</tr>
<tr>
<td>Izmir</td>
<td>4,462,056</td>
<td>3</td>
<td>4,395,653</td>
<td>937,607</td>
<td>53.4</td>
<td>1.18</td>
</tr>
<tr>
<td>Adana</td>
<td>2,274,106</td>
<td>7</td>
<td>3,531,441</td>
<td>709,075</td>
<td>51.3</td>
<td>0.99</td>
</tr>
<tr>
<td>Kayseri</td>
<td>1,441,523</td>
<td>15</td>
<td>2,633,349</td>
<td>479,849</td>
<td>51.7</td>
<td>1.06</td>
</tr>
<tr>
<td>Antalya</td>
<td>2,688,004</td>
<td>5</td>
<td>2,617,274</td>
<td>636,535</td>
<td>51.4</td>
<td>1.08</td>
</tr>
<tr>
<td>Konya</td>
<td>2,296,347</td>
<td>6</td>
<td>2,607,459</td>
<td>579,927</td>
<td>51.2</td>
<td>1.15</td>
</tr>
<tr>
<td>Bursa</td>
<td>3,194,720</td>
<td>4</td>
<td>2,408,817</td>
<td>525,016</td>
<td>52.3</td>
<td>1.18</td>
</tr>
<tr>
<td>Balikesir</td>
<td>1,257,590</td>
<td>17</td>
<td>2,322,111</td>
<td>425,171</td>
<td>55.7</td>
<td>1.13</td>
</tr>
<tr>
<td>Samsun</td>
<td>1,368,488</td>
<td>16</td>
<td>2,237,385</td>
<td>422,438</td>
<td>54.2</td>
<td>1.06</td>
</tr>
</tbody>
</table>

a2022 census data by the Turkish Statistical Institute (TurkStat).
bThe rank of cities in Turkey by total population count.
Table 4. Screening and monitoring encounters per module in January 2023.

<table>
<thead>
<tr>
<th>Module and encounter type</th>
<th>Monthly target</th>
<th>Number of encounters</th>
<th>Achievement rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hypertension</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td>3,868,662</td>
<td>789,699</td>
<td>20.4</td>
</tr>
<tr>
<td>Monitoring</td>
<td>4,795,117</td>
<td>709,004</td>
<td>14.8</td>
</tr>
<tr>
<td><strong>Obesity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td>4,849,306</td>
<td>1,083,414</td>
<td>22.3</td>
</tr>
<tr>
<td>Monitoring</td>
<td>53,770</td>
<td>51,512</td>
<td>95.8</td>
</tr>
<tr>
<td><strong>Diabetes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td>898,665</td>
<td>670,502</td>
<td>74.6</td>
</tr>
<tr>
<td>Monitoring</td>
<td>2,128,955</td>
<td>279,071</td>
<td>13.1</td>
</tr>
<tr>
<td><strong>CVD(^a) risk</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screening</td>
<td>742,634</td>
<td>262,419</td>
<td>35.3</td>
</tr>
<tr>
<td>Monitoring</td>
<td>1,488,004</td>
<td>589,523</td>
<td>39.6</td>
</tr>
<tr>
<td><strong>Older adult</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monitoring</td>
<td>720,928</td>
<td>73,697</td>
<td>10.2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>19,546,041</td>
<td>4,508,841</td>
<td>23.1</td>
</tr>
</tbody>
</table>

\(^a\)CVD: cardiovascular disease.

Table 5. Achievement rates of treatment goals.

<table>
<thead>
<tr>
<th>Treatment goal</th>
<th>Achievement rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic BP(^a)</td>
<td>88.8</td>
</tr>
<tr>
<td>Diastolic BP</td>
<td>94.2</td>
</tr>
<tr>
<td>Fasting glucose</td>
<td>52.0</td>
</tr>
<tr>
<td>HbA(_1c)(^b)</td>
<td>61.5</td>
</tr>
<tr>
<td>LDL(^c) cholesterol</td>
<td>14.8</td>
</tr>
<tr>
<td>HDL(^d) cholesterol</td>
<td>63.2</td>
</tr>
<tr>
<td>Triglyceride</td>
<td>52.6</td>
</tr>
<tr>
<td>Weight</td>
<td>5.6</td>
</tr>
<tr>
<td>BMI</td>
<td>6.3</td>
</tr>
<tr>
<td>Waist circumference</td>
<td>2.9</td>
</tr>
</tbody>
</table>

\(^a\)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 chronic disease management.
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, HbA1c 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.

References
17. von Tottelen M, Grinyer K, Arfa A, Traore L, Verdoy D, Keung SNLC, et al. An integrated 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]


32. What is FHIR? The Office of the National Coordinator for Health Information Technology. URL: https://www.healthit.gov/sites/default/files/2019-08/ONCFHIRFSWhatsFHIR.pdf [accessed 2023-11-02]


39. HL7 FHIR® based secure data repository, onFHIR.io. URL: https://onfhir.io/ [accessed 2023-11-02]


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]


Abbreviations

CDSS: clinical decision support service
DMP: Disease Management Platform
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
FMP: family medicine practitioner
HbA1c: hemoglobin A1c
HCP: health care professional
HL7: Health Level Seven
ICD-10: International Classification of Diseases, Tenth Revision
MOH: Ministry of Health
PHR: personal health record

©Mustafa Mahir Ulgu, Göke Banu Laleci Erturkmen, Mustafa Yuksel, Tuncay Namli, Şenan Postacı, Mert Gencturk, Yildiray Kabak, A Anil Sinaci, Suat Gonul, Asuman Dogac, Zübeyde Özkan Altunay, Banu Ekinci, Sahin Aydin, Suayip Birinci. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Ten Years of Experience With a Telemedicine Platform Dedicated to Health Care Personnel: Implementation Report

Claudio Azzolini\textsuperscript{1,2,3}, MD; Elias Premi\textsuperscript{3,4}, MD; Simone Donati\textsuperscript{3,5}, MD; Andrea Falco\textsuperscript{2,6}, MEng; Aldo Torreggiani\textsuperscript{7}, MEng; Francesco Sicurello\textsuperscript{3,8}, BSc; Andreina Baj\textsuperscript{4}, MD; Lorenzo Azzi\textsuperscript{9}, MD; Alessandro Orro\textsuperscript{2,3,8}, MEng; Giovanni Porta\textsuperscript{5}, MD; Giovanna Azzolini\textsuperscript{2}, BSc; Marco Sorrentino\textsuperscript{2,9}, JD; Paolo Melillo\textsuperscript{10}, MEng; Francesco Testa\textsuperscript{10}, MD; Francesca Simonelli\textsuperscript{10}, MD; Gianfranco Giardina\textsuperscript{11}, MEng; Umberto Paolucci\textsuperscript{12}, MEng

\textsuperscript{1}Advisory Council of e-Health and Telemedicine, University of Insubria of Varese-Como, Varese, Italy
\textsuperscript{2}TM9 Srl, Milan, Italy
\textsuperscript{3}Italian Association of Telemedicine and Medical Informatics, Milan, Italy
\textsuperscript{4}Department of Life Sciences and Biotechnologies, University of Insubria, Varese-Como, Italy
\textsuperscript{5}Department of Medicine and Surgery, University of Insubria, Varese-Como, Italy
\textsuperscript{6}Alfa Design Studio, Milan, Italy
\textsuperscript{7}T&C Srl, Milan, Italy
\textsuperscript{8}Institute of Biomedical Technologies, National Research Council, Milan, Italy
\textsuperscript{9}Bns Farm Law Firm, Milan, Italy
\textsuperscript{10}Multidisciplinary Department of Medical, Surgical and Dental Sciences, University of Campania Luigi Vanvitelli, Naples, Italy
\textsuperscript{11}DDay.it, Milan, Italy
\textsuperscript{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 Eumed 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.
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 Eumedea 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

https://medinform.jmir.org/2024/1/e42847

JMIR Med Inform 2024 | vol. 12 | e42847 | p.98
(page number not for citation purposes)
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.

<table>
<thead>
<tr>
<th>Module</th>
<th>Description</th>
<th>Module type</th>
<th>Purpose</th>
<th>Timeframe of project activity</th>
<th>Holder</th>
<th>Sponsor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Teleconsultation in retinal diseases [18]</td>
<td>Second opinions&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Feasibility of second opinions among physicians</td>
<td>1 Month (during 2011)</td>
<td>Insubria University, Varese-Como</td>
<td>Comed Research nonprofit association, Milan</td>
</tr>
<tr>
<td>2</td>
<td>Age-related maculopathy [19]</td>
<td>Group&lt;sup&gt;b&lt;/sup&gt; (10 locations)</td>
<td>Acceleration of anti-vascular endothelial growth factor therapy</td>
<td>19 Months (2011-2012)</td>
<td>T&amp;C Srl, Milan, Italy</td>
<td>Novartis Pharma SpA, Origgio, Italy</td>
</tr>
<tr>
<td>3</td>
<td>Retinal pathology samples and correlated genes [20]</td>
<td>Data&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Collection of data on gene expression and functionality</td>
<td>4 Months (2012-2013)</td>
<td>Insubria University, Varese-Como</td>
<td>Insubria University, Varese-Como</td>
</tr>
<tr>
<td>4</td>
<td>Epiretinal macular membrane [21]</td>
<td>Data</td>
<td>Collection of data on disease morphology and functionality</td>
<td>10 Months (2015-2016)</td>
<td>Insubria University, Varese-Como</td>
<td>Insubria University, Varese-Como</td>
</tr>
<tr>
<td>5</td>
<td>Inherited eye diseases</td>
<td>Data</td>
<td>Collection of data on genetic eye diseases</td>
<td>2017-present</td>
<td>Ophthalmological Unit II, University of Naples</td>
<td>Ophthalmological Unit II, University of Naples; Rome Foundation</td>
</tr>
<tr>
<td>7</td>
<td>Second opinions among resident physicians</td>
<td>Second opinions</td>
<td>Feasibility of second opinions in didactics</td>
<td>4 Months (during 2019)</td>
<td>Comed Research nonprofit association, Milan</td>
<td>Bayer Italy SpA, Milan</td>
</tr>
<tr>
<td>8</td>
<td>Instrumental data in multiple sclerosis</td>
<td>Group (2 locations)</td>
<td>Collection of multidisciplinary data on disease</td>
<td>2019-present</td>
<td>Neurological Unit, Insubria, University Varese-Como</td>
<td>Insubria University, Varese-Como</td>
</tr>
<tr>
<td>10</td>
<td>SARS-CoV-2 on throat and ocular surfaces [23]</td>
<td>Group (2 locations)</td>
<td>Search for SARS-CoV-2 in throat and tears in COVID-19 patients</td>
<td>1 Month (during 2020)</td>
<td>T&amp;C Srl, Milan, Italy</td>
<td>Insubria University, Varese-Como</td>
</tr>
<tr>
<td>11</td>
<td>Potential malignant oral lesions</td>
<td>Group (2 locations)</td>
<td>Collection of data on disease</td>
<td>2021-present</td>
<td>Orthodontics Unit, Insubria University, Varese-Como</td>
<td>Insubria University, Varese-Como</td>
</tr>
<tr>
<td>12</td>
<td>Maculopathies and anti-aging medicine</td>
<td>Data</td>
<td>Collection of data on diseases and follow-up</td>
<td>2022-present</td>
<td>Claude Boscher, MD</td>
<td>Claude Boscher, MD</td>
</tr>
</tbody>
</table>

<sup>a</sup>Second opinions: second opinions from health care professionals at the same or a different institution.

<sup>b</sup>Group: shared database used by health care professionals at more than one institution.

<sup>c</sup>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.

<table>
<thead>
<tr>
<th>Access to the network by computer or mobile devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Poor: score of 1</td>
</tr>
<tr>
<td>- Good: score of 2</td>
</tr>
<tr>
<td>- Very good: score of 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acceptability of the procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Poor: score of 1</td>
</tr>
<tr>
<td>- Good: score of 2</td>
</tr>
<tr>
<td>- Very good: score of 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medical efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Poor: score of 1</td>
</tr>
<tr>
<td>- Good: score of 2</td>
</tr>
<tr>
<td>- Very good: score of 3</td>
</tr>
</tbody>
</table>

**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.

### Score for AMD

<table>
<thead>
<tr>
<th>Score = 3</th>
<th>Score = 2</th>
<th>Score = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age:</td>
<td>Visual Acuity:</td>
<td>Distortion at Amsler:</td>
</tr>
<tr>
<td>&lt; 40 years</td>
<td>&gt;= 4/10</td>
<td>marked</td>
</tr>
<tr>
<td>60 - 80 years</td>
<td>4/10 - 2/10</td>
<td>slight</td>
</tr>
<tr>
<td>&gt; 80 years</td>
<td>&lt; 2/10</td>
<td>absent</td>
</tr>
<tr>
<td>2/10 - 10/10</td>
<td>2/10</td>
<td>2/10 - 10/10</td>
</tr>
</tbody>
</table>

Total Score: 12 - If score ≥ 8 send the patient to the center.

### 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).
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.

<table>
<thead>
<tr>
<th>Module</th>
<th>Description</th>
<th>Centers involved, n</th>
<th>HCPs(^a) involved, n</th>
<th>Participants whose data were inserted, n</th>
<th>Text/image fields for each EMR(^b), n (fields that were filled in, %)</th>
<th>Beneficial effects</th>
<th>Evaluation score(^c) (minimum positive score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Teleconsultation in retinal diseases [18]</td>
<td>1 Retina center, 17 territorial offices</td>
<td>18</td>
<td>52</td>
<td>30 (60)</td>
<td>Useful teleconsultation among doctors</td>
<td>109 (108)</td>
</tr>
<tr>
<td>2</td>
<td>Age-related maculopathy [19]</td>
<td>11 Retina centers</td>
<td>114</td>
<td>678</td>
<td>65 (85)</td>
<td>Improvements in patients’ functional final outcomes</td>
<td>803 (684)</td>
</tr>
<tr>
<td>3</td>
<td>Retinal pathology samples and correlated genes [20]</td>
<td>1 Ophthalmological center, 1 genetic center</td>
<td>11</td>
<td>12</td>
<td>65 (80)</td>
<td>Better understanding of molecular mechanisms</td>
<td>Not acquired</td>
</tr>
<tr>
<td>4</td>
<td>Epiretinal macular membrane [21]</td>
<td>2 Ophthalmological centers, 1 human anatomy center</td>
<td>11</td>
<td>28</td>
<td>25 (65)</td>
<td>Identification of ultramicroscopic features of membranes</td>
<td>80 (66)</td>
</tr>
<tr>
<td>5</td>
<td>Inherited eye diseases(^d)</td>
<td>1 Ophthalmological center, 1 genetic center</td>
<td>14</td>
<td>1145(^e)</td>
<td>480(^f) (20)</td>
<td>Increased knowledge of genetic eye diseases</td>
<td>In progress</td>
</tr>
<tr>
<td>6</td>
<td>Retinal dystrophy due to RPE65 gene mutation [22]</td>
<td>9 Retinal-genetic centers</td>
<td>28</td>
<td>60</td>
<td>260 (65)</td>
<td>Identification of suitable patients for therapy</td>
<td>200 (168)</td>
</tr>
<tr>
<td>7</td>
<td>Second opinions among resident physicians(^d)</td>
<td>4 University ophthalmological departments</td>
<td>19</td>
<td>110</td>
<td>12 (85)</td>
<td>Resident physicians’ learning accelerated</td>
<td>140 (114)</td>
</tr>
<tr>
<td>8</td>
<td>Instrumental data in multiple sclerosis(^d)</td>
<td>2 Neurological centers, 2 ophthalmological centers</td>
<td>6</td>
<td>58(^e)</td>
<td>450(^f) (18)</td>
<td>Recognition of the disease in the subclinical stage</td>
<td>In progress</td>
</tr>
<tr>
<td>9</td>
<td>Epidemiological data on COVID-19 in workers(^d)</td>
<td>2 Care offices at 2 airports</td>
<td>9</td>
<td>298</td>
<td>30 (90)</td>
<td>Collection of useful diagnostic data on COVID-19 and how the disease is transmitted</td>
<td>75 (54)</td>
</tr>
<tr>
<td>10</td>
<td>SARS-CoV-2 on throat and ocular surfaces [23]</td>
<td>14 Medical units</td>
<td>20</td>
<td>108</td>
<td>34 (95)</td>
<td>Increased knowledge of COVID-19</td>
<td>165 (120)</td>
</tr>
<tr>
<td>11</td>
<td>Potential malignant oral lesions</td>
<td>4 Medical units</td>
<td>6</td>
<td>15(^e)</td>
<td>50(^f) (68)</td>
<td>Better prevention and therapy</td>
<td>In progress</td>
</tr>
<tr>
<td>12</td>
<td>Maculopathies and anti-aging medicine</td>
<td>1 Retina center</td>
<td>6</td>
<td>10(^f)</td>
<td>110(^f) (58)</td>
<td>Significantly better care</td>
<td>In progress</td>
</tr>
</tbody>
</table>

\(^a\)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 Oggiouni (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
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

https://medinform.jmir.org/2024/1/e42847  JMIR Med Inform 2024 | vol. 12 | e42847 | p.108 (page number not for citation purposes)


27. Azzolini C, Donati S. The digital era: new horizons in medicine and rehabilitation. 2023 Presented at: XXII Infopoverty World Conference; December 1; New York, NY.


32. Azzolini C, Donati S. The digital era: new horizons in medicine and rehabilitation. 2023 Presented at: XXII National Congress of the Italian Association of Telemedicine and Medical Informatics; November 24-25; Rome, Italy.


Abbreviations

EU: European Union
EMR: electronic medical record
HCP: health care professional
mHealth: mobile health

https://medinform.jmir.org/2024/1/e42847
Learnings From Implementation of Technology-Enabled Mental Health Interventions in India: Implementation Report

Sudha Kallakuri1, MSc; Sridevi Gara1, BE; Mahesh Godi1, BE; Sandhya Kanaka Yatirajula1, PhD; Srilatha Paslawar1, MPH; Mercian Daniel1, PhD; David Peiris2,3, MBBS, MIPH, PhD; Pallab Kumar Maulik1,3,4,5,6, MSc, MD, PhD

1George Institute for Global Health, New Delhi, India
2George Institute for Global Health, Sydney, Australia
3Faculty of Medicine, University of New South Wales, Sydney, Australia
4Department of Brain Sciences, Imperial College London, London, United Kingdom
5Prasanna School of Public Health, Manipal Academy of Higher Education, Manipal, India
6George 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

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=MjMyNTQt=&Enc=&userName=CTRI/2018/08/015355. ARTEMIS trial: Clinical Trial Registry India CTRI/2022/02/040307; https://ctri.nic.in/Clinicaltrials/pmaindet2.php?EncHid=NDc4MTE=4&Enc=&userName=CTRI/2022/02/040307

(JMIR Med Inform 2024;12:e47504) doi:10.2196/47504

https://medinform.jmir.org/2024/12/e47504 JMIR Med Inform 2024 | vol. 12 | e47504 | p.111 (page number not for citation purposes)
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 among adults and adolescents in India.

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.

<table>
<thead>
<tr>
<th>App</th>
<th>Phase of the study</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMART MH (Pilot) and SMART MH trial focused on rural adults</td>
<td>Listing (household census data collection)</td>
<td>Data collectors</td>
</tr>
<tr>
<td>Screening app</td>
<td>Household screening for common mental disorders</td>
<td>ASHAŠb</td>
</tr>
<tr>
<td>Baseline data collection app</td>
<td>Baseline: collected data on different variables and stressors triggering anxiety/depression</td>
<td>Data collectors</td>
</tr>
<tr>
<td>Intervention (ASHA app)</td>
<td>Intervention: for regular follow up of adults at high risk of CMDs who sought care from the doctor or have yet to seek care</td>
<td>ASHAs</td>
</tr>
<tr>
<td>Intervention (doctor app)</td>
<td>Intervention: diagnosis and treatment for CMDs among adults</td>
<td>Primary care doctors</td>
</tr>
<tr>
<td>3M, 6M, and 12M app</td>
<td>Assessments at 3, 6, and 12 months of the intervention</td>
<td>Data collectors</td>
</tr>
<tr>
<td>ARTEMIS trial focused on adolescents</td>
<td>Listing (household census data collection)</td>
<td>Data collectors</td>
</tr>
<tr>
<td>Screening app</td>
<td>Household screening for common mental disorders</td>
<td>Data collectors</td>
</tr>
<tr>
<td>Baseline data collection app</td>
<td>Baseline: collected data on different variables and stressors triggering anxiety/depression</td>
<td>Data collectors</td>
</tr>
<tr>
<td>Intervention (ASHA app)</td>
<td>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</td>
<td>ASHAs</td>
</tr>
<tr>
<td>Intervention (doctor app)</td>
<td>Intervention: diagnosis and treatment for CMDs</td>
<td>Primary care doctors</td>
</tr>
<tr>
<td>3M, 6M, and 12M app</td>
<td>Assessments at 3, 6, and 12 months of the intervention</td>
<td>Data collectors</td>
</tr>
</tbody>
</table>

aSMART MH: Systematic Medical Appraisal and Referral Treatment (SMART) Mental Health.
bASHA: Accredited Social Health Activist.
cCMD: common mental disorder.
dARTEMIS: 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).
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).

Based on the scope of work (i.e., 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.
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.

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.

<table>
<thead>
<tr>
<th>Project</th>
<th>Study participants reached, n</th>
<th>ASHAs/CWVs, n</th>
<th>Doctors included, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMART MH Pilot (2014 to 2019)</td>
<td>50,000 adults</td>
<td>40</td>
<td>14</td>
</tr>
<tr>
<td>SMART MH Trial (2018 to 2022)</td>
<td>165,000 adults</td>
<td>175</td>
<td>50</td>
</tr>
<tr>
<td>ARTEMIS (2020-2024)</td>
<td>69,600 adolescents (10-19 years old)</td>
<td>104</td>
<td>27</td>
</tr>
</tbody>
</table>

ASHA: Accredited Social Health Activist.
CWV: community woman volunteer.
SMART MH: Systematic Medical Appraisal and Referral Treatment Mental Health.
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).

<table>
<thead>
<tr>
<th>Configuration of OpenMRS modules</th>
<th>Features for the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creation of a concept dictionary</td>
<td>Every data point to be used for the study was created as a concept and given a short name</td>
</tr>
<tr>
<td>Role management</td>
<td>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</td>
</tr>
<tr>
<td>User management</td>
<td>As per our project flow, the different users were allocated to each role, such as ASHAs, doctors, field staff/data collector, project manager, and administrator</td>
</tr>
<tr>
<td>Encounter management</td>
<td>Each entry into the tab for a specific user (i.e., 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</td>
</tr>
<tr>
<td>Managing encounter types</td>
<td>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</td>
</tr>
<tr>
<td>Manage observations</td>
<td>Each data point was considered as a separate observation</td>
</tr>
<tr>
<td>Managing persons</td>
<td>Demographic data for every app user (ASHA, doctor) or participant were stored as person details</td>
</tr>
<tr>
<td>Managing patients</td>
<td>In this feature, any additional personal identifiers/demographic details identified could be modified/configured</td>
</tr>
<tr>
<td>Cohort management</td>
<td>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 having a defined set of ASHAs, who in turn had a defined set of high-risk individuals</td>
</tr>
<tr>
<td>Multilocation data management</td>
<td>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 and ARTEMIS trials, which involved two different geographical locations.</td>
</tr>
</tbody>
</table>

aASHA: Accredited Social Health Activist.  
bPHC: primary health center.  
cSMH: Systematic Medical Appraisal and Referral Treatment Mental Health.  
dARTEMIS: 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.

<table>
<thead>
<tr>
<th>Issues that needed amendments</th>
<th>Solutions for the problems/issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>Development of descriptive analytics at the database level while implementing the SMH 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)</td>
</tr>
<tr>
<td>Monitoring an individual’s mental health status over time was not possible.</td>
<td>Analytics were developed to track the PHQ and GAD 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.</td>
</tr>
<tr>
<td>The performance of ASHAs could not be tracked well.</td>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
<td>The app is protected with multifactor authentication using a password and lock pin as an enhancement to the existing setup.</td>
</tr>
<tr>
<td>User interface and functioning of the app were not clear.</td>
<td>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.</td>
</tr>
<tr>
<td>Enabling online training during COVID-19.</td>
<td>Some of the training materials were embedded in the mobile apps to enable easier access for trainees using virtual modes during COVID-19.</td>
</tr>
<tr>
<td>Real-time monitoring of the activities of field staff was required to ensure increased data quality.</td>
<td>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.</td>
</tr>
<tr>
<td>Merging of two apps, namely household listing and participant screening, into one app.</td>
<td>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.</td>
</tr>
</tbody>
</table>

---

aSMH: Systematic Medical Appraisal and Referral Treatment Mental Health.
bPHQ9: Patient Health Questionnaire-9.
cGAD7: Generalized Anxiety Disorder-7.
dASHA: Accredited Social Health Activist.
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 ResIllience (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
15. OpenMRS. URL: https://openmrs.org/ [accessed 2024-01-22]


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
Value of Electronic Health Records Measured Using Financial and Clinical Outcomes: Quantitative Study

Shikha Modi¹, MBA, PhD; Sue S Feldman², RN, MEd, PhD; Eta S Berner², EdD; Benjamin Schooley³, PhD; Allen Johnston⁴, PhD

¹The University of Alabama in Huntsville, Huntsville, AL, United States
²The University of Alabama at Birmingham, Birmingham, AL, United States
³Brigham Young University, Provo, UT, United States
⁴Department 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 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;
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].
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 a 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).

**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.

**Textbox 1. Analytic models and their use.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Analytic Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>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.</td>
</tr>
<tr>
<td>Model 2</td>
<td>Determine the association between levels of EHR adoption and operating margin moderated by readmission rates with the operating margins from the same year.</td>
</tr>
<tr>
<td>Model 3</td>
<td>Determine the association between levels of EHR adoption and total margin moderated by LOS with the total margins from the same year.</td>
</tr>
<tr>
<td>Model 4</td>
<td>Determine the association between levels of EHR adoption and total margin moderated by readmission rates with the total margins from the same year.</td>
</tr>
<tr>
<td>Model 5</td>
<td>Determine the association between levels of EHR adoption and operating margin moderated by LOS with a 1-year lag in the operating margins.</td>
</tr>
<tr>
<td>Model 6</td>
<td>Determine the association between levels of EHR adoption and operating margin moderated by LOS with a 2-year lag in the operating margins.</td>
</tr>
<tr>
<td>Model 7</td>
<td>Determine the association between levels of EHR adoption and operating margin moderated by LOS with a 1-year lag in the operating margins.</td>
</tr>
<tr>
<td>Model 8</td>
<td>Determine the association between levels of EHR adoption and operating margin moderated by LOS with a 2-year lag in the operating margins.</td>
</tr>
<tr>
<td>Model 9</td>
<td>Determine the association between levels of EHR adoption and total margin moderated by LOS with a 1-year lag in the total margins.</td>
</tr>
<tr>
<td>Model 10</td>
<td>Determine the association between levels of EHR adoption and total margin moderated by LOS with a 2-year lag in the total margins.</td>
</tr>
<tr>
<td>Model 11</td>
<td>Determine the association between levels of EHR adoption and total margin moderated by readmission rates with a 1-year lag in the total margins.</td>
</tr>
<tr>
<td>Model 12</td>
<td>Determine the association between levels of EHR adoption and total margin moderated by readmission rates with a 2-year lag in the total margins.</td>
</tr>
</tbody>
</table>

In this equation, $y_{it}$ is the dependent variable (financial or clinical outcomes), $i =$ hospital, and $t =$ time. $\beta_1$ is the coefficient for the main independent variable (levels of EHR adoption), $X_{it1}$. $\beta_2$ is the coefficient for the moderator variable (clinical outcomes), $X_{it2}$. $\beta_3$ 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). $\alpha_i$ is the unknown intercept for a vector of hospitals. And $\mu_{it}$ 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.
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).

<table>
<thead>
<tr>
<th>Variables</th>
<th>2014 (n=1420)</th>
<th>2015 (n=1453)</th>
<th>2016 (n=1393)</th>
<th>2017 (n=1412)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levels of EHR&lt;sup&gt;a&lt;/sup&gt; adoption, mean (SD)</td>
<td>0.871 (0.127)</td>
<td>0.890 (0.121)</td>
<td>0.899 (0.127)</td>
<td>0.917 (0.102)</td>
</tr>
<tr>
<td>Operating margin, mean (SD)</td>
<td>0.070 (0.122)</td>
<td>0.065 (0.132)</td>
<td>0.06 (0.140)</td>
<td>0.057 (0.136)</td>
</tr>
<tr>
<td>Total margin, mean (SD)</td>
<td>1.014 (2.847)</td>
<td>0.005 (26.4)</td>
<td>1.136 (7.217)</td>
<td>0.951 (1.129)</td>
</tr>
<tr>
<td>Average length of stay (days), mean (SD)</td>
<td>3.911 (1.134)</td>
<td>3.881 (0.954)</td>
<td>7.87 (153.3)</td>
<td>7.945 (160.4)</td>
</tr>
<tr>
<td>Readmission rate, mean (SD)</td>
<td>0.998 (0.003)</td>
<td>0.995 (0.006)</td>
<td>0.995 (0.006)</td>
<td>0.994 (0.007)</td>
</tr>
<tr>
<td>Market competition (HHI&lt;sup&gt;b&lt;/sup&gt;) in terms of discharges, mean (SD)</td>
<td>0.101 (0.199)</td>
<td>0.086 (0.157)</td>
<td>0.088 (0.172)</td>
<td>0.098 (0.193)</td>
</tr>
<tr>
<td>Medicare percentage, mean (SD)</td>
<td>0.512 (0.140)</td>
<td>0.518 (0.128)</td>
<td>0.518 (0.130)</td>
<td>0.512 (0.124)</td>
</tr>
<tr>
<td>Medicaid percentage, mean (SD)</td>
<td>0.197 (0.120)</td>
<td>0.202 (0.115)</td>
<td>0.203 (0.114)</td>
<td>0.204 (0.112)</td>
</tr>
<tr>
<td>Beds (n), mean (SD)</td>
<td>257 (231)</td>
<td>256 (229)</td>
<td>254 (232)</td>
<td>255 (236)</td>
</tr>
<tr>
<td>Ownership, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nongovernment not-for-profit</td>
<td>1105 (77.76 )</td>
<td>1145 (78.8 )</td>
<td>1177 (78.31)</td>
<td>1198 (78.82)</td>
</tr>
<tr>
<td>Investor-owned for-profit</td>
<td>294 (20.69)</td>
<td>295 (20.30)</td>
<td>311 (20.69)</td>
<td>305 (20.07)</td>
</tr>
<tr>
<td>Government nonfederal</td>
<td>22 (1.55)</td>
<td>13 (0.89)</td>
<td>15 (1)</td>
<td>17 (1.12)</td>
</tr>
<tr>
<td>Affiliation, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>584 (47.29)</td>
<td>660 (51.36)</td>
<td>687 (51.58)</td>
<td>731 (56.67)</td>
</tr>
<tr>
<td>No</td>
<td>651 (52.71)</td>
<td>625 (48.64)</td>
<td>645 (48.42)</td>
<td>559 (43.33)</td>
</tr>
<tr>
<td>Teaching status, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>560 (39.41)</td>
<td>569 (39.16)</td>
<td>595 (39.59)</td>
<td>599 (39.41)</td>
</tr>
<tr>
<td>No</td>
<td>861 (60.59)</td>
<td>884 (60.84)</td>
<td>908 (60.41)</td>
<td>921 (60.59)</td>
</tr>
</tbody>
</table>

<sup>a</sup>EHR: electronic health record.

<sup>b</sup>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.

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>Independent variables: levels of EHR(^a) adoption (correlation coefficients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating margin</td>
<td>0.0978(^b)</td>
</tr>
<tr>
<td>Total margin</td>
<td>-0.0142</td>
</tr>
<tr>
<td>Average length of stay</td>
<td>0.0039</td>
</tr>
<tr>
<td>Readmission rate</td>
<td>-0.0321(^b)</td>
</tr>
</tbody>
</table>

\(^a\)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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Levels of EHR adoption</td>
<td>-0.020</td>
<td>5.335(^f)</td>
<td>-4.961</td>
<td>415.2</td>
</tr>
<tr>
<td><strong>Dependent variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average LOS</td>
<td>0.000</td>
<td>N/A(^e)</td>
<td>-0.002</td>
<td>N/A</td>
</tr>
<tr>
<td>RR</td>
<td>N/A</td>
<td>4.375(^h)</td>
<td>N/A</td>
<td>431.6</td>
</tr>
<tr>
<td>Levels of EHR adoption and average LOS</td>
<td>-0.000</td>
<td>N/A</td>
<td>0.001</td>
<td>N/A</td>
</tr>
<tr>
<td>Levels of EHR adoption and RR</td>
<td>N/A</td>
<td>-5.384(^f)</td>
<td>N/A</td>
<td>-422.3</td>
</tr>
<tr>
<td><strong>Control variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market competition (HHI(^i))</td>
<td>0.082</td>
<td>0.078</td>
<td>3.148</td>
<td>2.959</td>
</tr>
<tr>
<td>Medicare percentage</td>
<td>-0.009</td>
<td>-0.013</td>
<td>0.699</td>
<td>0.937</td>
</tr>
<tr>
<td>Medicaid percentage</td>
<td>-0.026</td>
<td>-0.026</td>
<td>1.211</td>
<td>1.343</td>
</tr>
<tr>
<td><strong>Years</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014 Reference</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>2015 0.005</td>
<td>-0.007(^f)</td>
<td>-1.001</td>
<td>-0.848</td>
<td></td>
</tr>
<tr>
<td>2016 0.008</td>
<td>-0.009(^f)</td>
<td>0.388</td>
<td>0.569</td>
<td></td>
</tr>
<tr>
<td>2017 0.008</td>
<td>-0.011(^j)</td>
<td>0.243</td>
<td>0.460</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)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.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 5</th>
<th>Model 6</th>
<th>Model 7</th>
<th>Model 8</th>
<th>Model 9</th>
<th>Model 10</th>
<th>Model 11</th>
<th>Model 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>OM&lt;sup&gt;a&lt;/sup&gt;–LOS&lt;sup&gt;b&lt;/sup&gt;–levels of EHR&lt;sup&gt;c&lt;/sup&gt; adoption with 1-year lag (Prob&gt;F=0.0761)</td>
<td>0.022</td>
<td>0.004</td>
<td>1.681</td>
<td>2.229</td>
<td>1.564</td>
<td>—1.547</td>
<td>—164.0</td>
<td>268.8</td>
</tr>
<tr>
<td>Average LOS</td>
<td>0.000</td>
<td>—9.46e–06</td>
<td>N/A&lt;sup&gt;f&lt;/sup&gt;</td>
<td>N/A</td>
<td>0.001</td>
<td>—0.012</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>RR</td>
<td>N/A</td>
<td>N/A</td>
<td>0.818</td>
<td>2.192</td>
<td>N/A</td>
<td>N/A</td>
<td>—186.4</td>
<td>169.8</td>
</tr>
<tr>
<td>Levels of EHR adoption and average LOS</td>
<td>—0.000</td>
<td>—4.26e–06</td>
<td>N/A</td>
<td>N/A</td>
<td>—0.002</td>
<td>0.013</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Levels of EHR adoption and RR</td>
<td>N/A</td>
<td>N/A</td>
<td>—1.663</td>
<td>—2.232</td>
<td>N/A</td>
<td>N/A</td>
<td>166.2</td>
<td>—271.7</td>
</tr>
<tr>
<td>Market competition (HHI&lt;sup&gt;g&lt;/sup&gt;)</td>
<td>0.219&lt;sup&gt;h&lt;/sup&gt;</td>
<td>—0.068</td>
<td>0.223&lt;sup&gt;h&lt;/sup&gt;</td>
<td>—0.068</td>
<td>3.610</td>
<td>—3.275</td>
<td>3.749</td>
<td>—3.103</td>
</tr>
<tr>
<td>Medicare percentage</td>
<td>0.063&lt;sup&gt;i&lt;/sup&gt;</td>
<td>0.024</td>
<td>0.068&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.030</td>
<td>—2.419</td>
<td>0.523</td>
<td>—2.416</td>
<td>0.783</td>
</tr>
<tr>
<td>Medicaid percentage</td>
<td>0.018</td>
<td>0.891&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.018</td>
<td>0.092&lt;sup&gt;h&lt;/sup&gt;</td>
<td>1.742</td>
<td>—2.822</td>
<td>1.741</td>
<td>—2.838</td>
</tr>
<tr>
<td>Years</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
<td>Reference</td>
</tr>
<tr>
<td>2014</td>
<td>0.014&lt;sup&gt;h&lt;/sup&gt;</td>
<td>—0.006</td>
<td>0.014&lt;sup&gt;h&lt;/sup&gt;</td>
<td>—0.006</td>
<td>—1.057</td>
<td>—0.714</td>
<td>—1.178</td>
<td>—0.910</td>
</tr>
<tr>
<td>2015</td>
<td>0.011&lt;sup&gt;h&lt;/sup&gt;</td>
<td>—0.008&lt;sup&gt;i&lt;/sup&gt;</td>
<td>0.009</td>
<td>—0.008&lt;sup&gt;i&lt;/sup&gt;</td>
<td>0.141</td>
<td>0.482</td>
<td>0.048</td>
<td>0.323</td>
</tr>
<tr>
<td>2016</td>
<td>0.010&lt;sup&gt;h&lt;/sup&gt;</td>
<td>—0.018&lt;sup&gt;i&lt;/sup&gt;</td>
<td>0.009&lt;sup&gt;i&lt;/sup&gt;</td>
<td>—0.018&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.008</td>
<td>0.508</td>
<td>—0.126</td>
<td>0.229</td>
</tr>
</tbody>
</table>

<sup>a</sup>OM: operating margin.

<sup>b</sup>LOS: length of stay.

<sup>c</sup>EHR: electronic health record.

<sup>d</sup>RR: readmission rate.

<sup>e</sup>TM: total margin.

<sup>f</sup>N/A: not applicable.

<sup>g</sup>HHI: Herfindahl-Hirschman Index.

<sup>h</sup>P<.05.

<sup>i</sup>P<.10.

<sup>j</sup>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.

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 LOS 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


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

©Shikha Modi, Sue S Feldman, Eta S Berner, Benjamin Schooley, Allen Johnston. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 24.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Additional Value From Free-Text Diagnoses in Electronic Health Records: Hybrid Dictionary and Machine Learning Classification Study

Tarun Mehra¹, MD; Tobias Wekhof², PhD; Dagmar Iris Keller³,⁴, MD

¹Department for Medical Oncology and Hematology, University Hospital of Zurich, Zurich, Switzerland
²Center of Economic Research, ETH Zurich, Zurich, Switzerland
³Faculty of Medicine, University of Zurich, Zurich, Switzerland
⁴Emergency Department, University Hospital of Zurich, Zurich, Switzerland

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—inferring the possibility of in-hospital mortality, unplanned readmission, 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.

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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y), mean (SD)</td>
<td>46.5 (19.7)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>23,782 (45.54)</td>
</tr>
<tr>
<td>Emergency Severity Index score (out of 5), mean (SD)</td>
<td>3.3 (0.6)</td>
</tr>
<tr>
<td>Fast track, n (%)</td>
<td>8264 (15.82)</td>
</tr>
<tr>
<td>Number of patients in the emergency department, mean (SD)</td>
<td>19.8 (8.3)</td>
</tr>
<tr>
<td>Early shift, n (%)</td>
<td>21,644 (41.45)</td>
</tr>
<tr>
<td>With emergency medical service, n (%)</td>
<td>9020 (17.27)</td>
</tr>
<tr>
<td>With police, n (%)</td>
<td>188 (0.36)</td>
</tr>
<tr>
<td>Accident, n (%)</td>
<td>16,845 (32.26)</td>
</tr>
<tr>
<td>Inpatient, n (%)</td>
<td>14,112 (27.02)</td>
</tr>
<tr>
<td>Night shift, n (%)</td>
<td>7915 (15.16)</td>
</tr>
<tr>
<td>Late shift, n (%)</td>
<td>22,663 (43.4)</td>
</tr>
</tbody>
</table>

aThe total sample contains patient visits for the period from May 28, 2019, to October 31, 2021.
Figure 2. Data structure.

| January 1, 2014 to October 31, 2021 |
|---|---|---|---|---|---|---|---|
| 35,356 | 36,612 | 38,526 | 38,296 | 39,432 | 40,227 | 34,232 | 30,614 |

Total patient visits: **293,298**

Exclude 36,969 patients with no text entry

Total patient visits: **256,329**

May 28, 2019 to October 31, 2021

Drop-down menu for triage

Total patient visits: **52,222**

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 (e.g., 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.
Table 2. Spherical feature annotation and increase in topic share (n=52,222)\textsuperscript{a}.

<table>
<thead>
<tr>
<th>Clinical topic NLP\textsuperscript{b}</th>
<th>Records tagged initially, n (%)</th>
<th>Records tagged NLP-augmented, n (%)</th>
<th>Increase in tagged patient records, n (%)\textsuperscript{c}</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19</td>
<td>375 (0.72)</td>
<td>405 (0.78)</td>
<td>30 (8)</td>
</tr>
<tr>
<td>General symptom</td>
<td>6401 (12.26)</td>
<td>6867 (13.15)</td>
<td>466 (7.28)</td>
</tr>
<tr>
<td>General administration</td>
<td>315 (0.6)</td>
<td>1217 (2.33)</td>
<td>902 (286.35)</td>
</tr>
<tr>
<td>Systemic clinical</td>
<td>3219 (6.16)</td>
<td>3519 (6.74)</td>
<td>300 (9.32)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>3421 (6.55)</td>
<td>4159 (7.96)</td>
<td>738 (21.57)</td>
</tr>
<tr>
<td>Respiratory</td>
<td>4040 (6.55)</td>
<td>4159 (7.96)</td>
<td>738 (21.57)</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>2683 (5.14)</td>
<td>5219 (9.99)</td>
<td>2536 (94.52)</td>
</tr>
<tr>
<td>Neurological</td>
<td>414 (7.93)</td>
<td>4485 (8.59)</td>
<td>345 (8.33)</td>
</tr>
<tr>
<td>Eye; ear, nose, and throat; and derma</td>
<td>1818 (3.48)</td>
<td>2061 (3.95)</td>
<td>243 (13.37)</td>
</tr>
<tr>
<td>Gynecology and urology</td>
<td>2712 (5.19)</td>
<td>3004 (5.75)</td>
<td>292 (10.77)</td>
</tr>
<tr>
<td>Trauma</td>
<td>16,516 (31.63)</td>
<td>17,389 (33.3)</td>
<td>873 (5.29)</td>
</tr>
<tr>
<td>General psychiatric</td>
<td>1989 (3.81)</td>
<td>2627 (5.03)</td>
<td>638 (32.08)</td>
</tr>
<tr>
<td>No tag</td>
<td>14,141 (27.08)</td>
<td>10,569 (20.24)</td>
<td>–3572 (–25.26)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}This table presents the distribution of the diagnosis topics obtained with the NLP-based text annotation before and after the spherical feature annotation.  
\textsuperscript{b}NLP: natural language processing.  
\textsuperscript{c}Percent of initially recorded tags.
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>
<thead>
<tr>
<th>Cluster and subcluster detail</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19</td>
<td>401 (0.77)</td>
</tr>
<tr>
<td><strong>General symptoms</strong></td>
<td></td>
</tr>
<tr>
<td>Fever</td>
<td>2440 (4.67)</td>
</tr>
<tr>
<td>Pain</td>
<td>4505 (8.63)</td>
</tr>
<tr>
<td>General weakness</td>
<td>80 (0.15)</td>
</tr>
<tr>
<td>Back pain</td>
<td>438 (0.84)</td>
</tr>
<tr>
<td><strong>General organizational</strong></td>
<td>1217 (2.33)</td>
</tr>
<tr>
<td>Follow-up and prescription</td>
<td>1217 (2.33)</td>
</tr>
<tr>
<td><strong>Systemic</strong></td>
<td>3519 (6.74)</td>
</tr>
<tr>
<td>Infection not otherwise specified</td>
<td>1239 (2.37)</td>
</tr>
<tr>
<td>Sepsis</td>
<td>125 (0.24)</td>
</tr>
<tr>
<td>Anaphylaxia and allergy</td>
<td>261 (0.5)</td>
</tr>
<tr>
<td>Cancer</td>
<td>1688 (3.23)</td>
</tr>
<tr>
<td>Transplantation</td>
<td>227 (0.43)</td>
</tr>
<tr>
<td>Glycemia</td>
<td>138 (0.26)</td>
</tr>
<tr>
<td><strong>Gastrointestinal</strong></td>
<td>4147 (7.94)</td>
</tr>
<tr>
<td>Gastrointestinal bleeding</td>
<td>522 (1)</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>1879 (3.6)</td>
</tr>
<tr>
<td>Diarrhea, vomiting, and nausea</td>
<td>2248 (4.3)</td>
</tr>
<tr>
<td><strong>Respiratory</strong></td>
<td>4311 (8.26)</td>
</tr>
<tr>
<td>Upper airway</td>
<td>1592 (3.05)</td>
</tr>
<tr>
<td>Lower airway</td>
<td>1934 (3.7)</td>
</tr>
<tr>
<td>Influenza</td>
<td>440 (0.84)</td>
</tr>
<tr>
<td>Dyspnea</td>
<td>2197 (4.21)</td>
</tr>
<tr>
<td><strong>Cardiovascular</strong></td>
<td>5211 (9.98)</td>
</tr>
<tr>
<td>Chest pain</td>
<td>3569 (6.83)</td>
</tr>
<tr>
<td>Palpitations and arrhythmia</td>
<td>518 (0.99)</td>
</tr>
<tr>
<td>Pulmonary embolism</td>
<td>281 (0.54)</td>
</tr>
<tr>
<td>Deep venous thrombosis</td>
<td>528 (1.01)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>394 (0.75)</td>
</tr>
<tr>
<td><strong>Neurological</strong></td>
<td>4466 (8.55)</td>
</tr>
<tr>
<td>Headache</td>
<td>1189 (2.28)</td>
</tr>
<tr>
<td>Neurological</td>
<td>1737 (3.33)</td>
</tr>
<tr>
<td>Vigilance and disorientation</td>
<td>191 (0.37)</td>
</tr>
<tr>
<td>Dementia</td>
<td>24 (0.05)</td>
</tr>
<tr>
<td>Syncope</td>
<td>453 (0.87)</td>
</tr>
<tr>
<td>Vertigo and dizziness</td>
<td>934 (1.79)</td>
</tr>
<tr>
<td>Convulsion</td>
<td>226 (0.43)</td>
</tr>
<tr>
<td><strong>Eye; ear, nose, and throat; and skin</strong></td>
<td>2061 (3.95)</td>
</tr>
<tr>
<td>Epistaxis</td>
<td>58 (0.11)</td>
</tr>
<tr>
<td>Eye symptoms</td>
<td>703 (1.35)</td>
</tr>
<tr>
<td>Hearing and auricular</td>
<td>18 (0.03)</td>
</tr>
<tr>
<td>Cluster and subcluster detail</td>
<td>Values, n (%)</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Skin</td>
<td>1311 (2.51)</td>
</tr>
<tr>
<td>Urological and gynecological</td>
<td></td>
</tr>
<tr>
<td>Urological and kidney</td>
<td>2973 (5.69)</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>34 (0.07)</td>
</tr>
<tr>
<td>Trauma</td>
<td>17,302 (33.13)</td>
</tr>
<tr>
<td>Wound</td>
<td>5163 (9.89)</td>
</tr>
<tr>
<td>Fracture and luxation</td>
<td>5375 (10.29)</td>
</tr>
<tr>
<td>Trauma and head</td>
<td>2171 (4.16)</td>
</tr>
<tr>
<td>Burns</td>
<td>141 (0.27)</td>
</tr>
<tr>
<td>Fall</td>
<td>729 (1.4)</td>
</tr>
<tr>
<td>Trauma not otherwise specified</td>
<td>9278 (17.77)</td>
</tr>
<tr>
<td>Bleeding not otherwise specified</td>
<td>986 (1.89)</td>
</tr>
<tr>
<td>Collision</td>
<td>1250 (2.39)</td>
</tr>
<tr>
<td>Traffic</td>
<td>314 (0.6)</td>
</tr>
<tr>
<td>Psychiatric</td>
<td>2625 (5.03)</td>
</tr>
<tr>
<td>Intoxication</td>
<td>1146 (2.19)</td>
</tr>
<tr>
<td>Psychiatric</td>
<td>851 (1.63)</td>
</tr>
<tr>
<td>Fear</td>
<td>725 (1.39)</td>
</tr>
<tr>
<td>Severity</td>
<td></td>
</tr>
<tr>
<td>Nonsevere</td>
<td>113 (0.22)</td>
</tr>
<tr>
<td>Severe</td>
<td>235 (0.45)</td>
</tr>
<tr>
<td>Chronic</td>
<td>55 (0.11)</td>
</tr>
<tr>
<td>Acute</td>
<td>232 (0.44)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>LS(^b), (n)</th>
<th>NLP(^c), (n)</th>
<th>Correlation (r)(^d)</th>
<th>Consistency(^e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVID-19</td>
<td>7623</td>
<td>401</td>
<td>0.18</td>
<td>0.05</td>
</tr>
<tr>
<td>General symptom</td>
<td>7993</td>
<td>6852</td>
<td>−0.04</td>
<td>0.10</td>
</tr>
<tr>
<td>General administration</td>
<td>642</td>
<td>1217</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>Systemic clinical</td>
<td>1983</td>
<td>3519</td>
<td>0.12</td>
<td>0.22</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>4063</td>
<td>4147</td>
<td>0.41</td>
<td>0.46</td>
</tr>
<tr>
<td>Respiratory</td>
<td>872</td>
<td>4311</td>
<td>0.17</td>
<td>0.44</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>2245</td>
<td>5211</td>
<td>0.28</td>
<td>0.49</td>
</tr>
<tr>
<td>Neurological</td>
<td>5123</td>
<td>4466</td>
<td>0.44</td>
<td>0.46</td>
</tr>
<tr>
<td>Eye; ear, nose, and throat; and derma</td>
<td>1041</td>
<td>2061</td>
<td>0.26</td>
<td>0.39</td>
</tr>
<tr>
<td>Gynecology and urology</td>
<td>1206</td>
<td>3004</td>
<td>0.40</td>
<td>0.67</td>
</tr>
<tr>
<td>Trauma</td>
<td>12,337</td>
<td>17,302</td>
<td>0.54</td>
<td>0.79</td>
</tr>
<tr>
<td>General psychiatric</td>
<td>1610</td>
<td>2625</td>
<td>0.60</td>
<td>0.78</td>
</tr>
<tr>
<td>No tag</td>
<td>5994</td>
<td>10,644</td>
<td>0.07</td>
<td>0.28</td>
</tr>
</tbody>
</table>

\(^a\)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, 8,950 (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:

Model 1: \(Y_i = \alpha + \beta X_i + \gamma Z_i + \epsilon_i\) (1)

Model 2: \(Y_i = \alpha + \beta X_i + \delta W_i + \epsilon_i\) (2)

Model 3: \(Y_i = \alpha + \beta X_i + \gamma Z_i + \delta W_i + \epsilon_i\) (3)

where \(Y_i\) is either the ESI or inpatient indicator variable for patient visit \(i\), \(\alpha\) 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 \(\epsilon_i\) 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”\(^a\).

<table>
<thead>
<tr>
<th>Name of cluster(^b)</th>
<th>Model 1(^c), regression coefficient (SE)</th>
<th>Model 2(^c), regression coefficient (SE)</th>
<th>Model 3 including both measures(^d), regression coefficient (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP(^e) cluster: COVID-19</td>
<td>0.048(^f) (0.019)</td>
<td>N/A(^g)</td>
<td>(-0.022) (0.022)</td>
</tr>
<tr>
<td>Chief complaint cluster: COVID-19</td>
<td>N/A</td>
<td>0.127(^f) (0.007)</td>
<td>0.133(^b) (0.008)</td>
</tr>
<tr>
<td>NLP cluster: general symptoms</td>
<td>0.011(^f) (0.005)</td>
<td>N/A</td>
<td>(-0.019) (0.005)</td>
</tr>
<tr>
<td>Chief complaint cluster: general symptoms</td>
<td>N/A</td>
<td>(-0.002) (0.007)</td>
<td>0.000 (0.007)</td>
</tr>
<tr>
<td>NLP cluster: general organizational</td>
<td>(-0.004) (0.011)</td>
<td>N/A</td>
<td>0.006 (0.011)</td>
</tr>
<tr>
<td>Chief complaint cluster: general organizational</td>
<td>N/A</td>
<td>(-0.062) (0.016)</td>
<td>(-0.052) (0.016)</td>
</tr>
<tr>
<td>NLP cluster: systemic</td>
<td>0.117(^b) (0.007)</td>
<td>N/A</td>
<td>0.101(^b) (0.007)</td>
</tr>
<tr>
<td>Chief complaint cluster: systemic</td>
<td>N/A</td>
<td>0.118(^b) (0.010)</td>
<td>0.104(^b) (0.010)</td>
</tr>
<tr>
<td>NLP cluster: gastrointestinal</td>
<td>0.071(^b) (0.006)</td>
<td>N/A</td>
<td>0.049(^b) (0.008)</td>
</tr>
<tr>
<td>Chief complaint cluster: gastrointestinal</td>
<td>N/A</td>
<td>0.083(^b) (0.008)</td>
<td>0.059(^b) (0.008)</td>
</tr>
<tr>
<td>NLP cluster: respiratory</td>
<td>0.063(^b) (0.007)</td>
<td>N/A</td>
<td>(-0.017) (0.008)</td>
</tr>
<tr>
<td>Chief complaint cluster: respiratory</td>
<td>N/A</td>
<td>0.133(^b) (0.014)</td>
<td>0.126(^f) (0.014)</td>
</tr>
<tr>
<td>NLP cluster: cardiovascular</td>
<td>(-0.020) (0.006)</td>
<td>N/A</td>
<td>(-0.009) (0.006)</td>
</tr>
<tr>
<td>Chief complaint cluster: cardiovascular</td>
<td>N/A</td>
<td>(-0.038) (0.010)</td>
<td>(-0.031) (0.010)</td>
</tr>
<tr>
<td>NLP cluster: neurological</td>
<td>(-0.046) (0.007)</td>
<td>N/A</td>
<td>(-0.045) (0.007)</td>
</tr>
<tr>
<td>Chief complaint cluster: neurological</td>
<td>N/A</td>
<td>(-0.058) (0.009)</td>
<td>(-0.048) (0.009)</td>
</tr>
<tr>
<td>NLP cluster: eye, ENT(^i), or skin</td>
<td>(-0.055) (0.009)</td>
<td>N/A</td>
<td>(-0.044) (0.009)</td>
</tr>
<tr>
<td>Chief complaint cluster: eye, ENT, or skin</td>
<td>N/A</td>
<td>(-0.128) (0.013)</td>
<td>(-0.112) (0.013)</td>
</tr>
<tr>
<td>NLP cluster: urological or gynecological</td>
<td>(-0.015) (0.008)</td>
<td>N/A</td>
<td>(-0.004) (0.008)</td>
</tr>
<tr>
<td>Chief complaint cluster: urological or gynecological</td>
<td>N/A</td>
<td>(-0.033) (0.012)</td>
<td>(-0.036) (0.013)</td>
</tr>
<tr>
<td>NLP cluster: trauma</td>
<td>(-0.041) (0.005)</td>
<td>N/A</td>
<td>(-0.038) (0.005)</td>
</tr>
<tr>
<td>Chief complaint cluster: trauma</td>
<td>N/A</td>
<td>0.011 (0.007)</td>
<td>0.020(^b) (0.007)</td>
</tr>
<tr>
<td>NLP cluster: psychiatric</td>
<td>(-0.079) (0.009)</td>
<td>N/A</td>
<td>(-0.053) (0.010)</td>
</tr>
<tr>
<td>Chief complaint cluster: psychiatric</td>
<td>N/A</td>
<td>(-0.068) (0.013)</td>
<td>(-0.039) (0.014)</td>
</tr>
</tbody>
</table>

\(^a\)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”

<table>
<thead>
<tr>
<th>Name of cluster</th>
<th>Model 1, regression coefficient (SE)</th>
<th>Model 2, regression coefficient (SE)</th>
<th>Model 3 including both measures, regression coefficient (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP cluster: COVID-19</td>
<td>0.079 (0.019)</td>
<td>N/A</td>
<td>0.023 (0.019)</td>
</tr>
<tr>
<td>Chief complaint cluster: COVID-19</td>
<td>N/A</td>
<td>0.214 (0.007)</td>
<td>0.172 (0.007)</td>
</tr>
<tr>
<td>NLP cluster: general symptoms</td>
<td>0.036 (0.005)</td>
<td>N/A</td>
<td>−0.023 (0.005)</td>
</tr>
<tr>
<td>Chief complaint cluster: general symptoms</td>
<td>N/A</td>
<td>−0.142 (0.007)</td>
<td>0.127 (0.007)</td>
</tr>
<tr>
<td>NLP cluster: general organizational</td>
<td>−0.050 (0.011)</td>
<td>N/A</td>
<td>−0.044 (0.011)</td>
</tr>
<tr>
<td>Chief complaint cluster: general organizational</td>
<td>N/A</td>
<td>0.308 (0.016)</td>
<td>0.352 (0.016)</td>
</tr>
<tr>
<td>NLP cluster: systemic</td>
<td>0.076 (0.007)</td>
<td>N/A</td>
<td>0.093 (0.007)</td>
</tr>
<tr>
<td>Chief complaint cluster: systemic</td>
<td>N/A</td>
<td>0.009 (0.010)</td>
<td>0.009 (0.010)</td>
</tr>
<tr>
<td>NLP cluster: gastrointestinal</td>
<td>0.192 (0.006)</td>
<td>N/A</td>
<td>0.088 (0.007)</td>
</tr>
<tr>
<td>Chief complaint cluster: gastrointestinal</td>
<td>N/A</td>
<td>0.305 (0.008)</td>
<td>0.262 (0.008)</td>
</tr>
<tr>
<td>NLP cluster: respiratory</td>
<td>0.114 (0.007)</td>
<td>N/A</td>
<td>0.053 (0.007)</td>
</tr>
<tr>
<td>Chief complaint cluster: respiratory</td>
<td>N/A</td>
<td>0.121 (0.014)</td>
<td>0.088 (0.014)</td>
</tr>
<tr>
<td>NLP cluster: cardiovascular</td>
<td>0.050 (0.006)</td>
<td>N/A</td>
<td>0.030 (0.006)</td>
</tr>
<tr>
<td>Chief complaint cluster: cardiovascular</td>
<td>N/A</td>
<td>0.205 (0.009)</td>
<td>0.197 (0.010)</td>
</tr>
<tr>
<td>NLP cluster: neurological</td>
<td>−0.015 (0.007)</td>
<td>N/A</td>
<td>−0.002 (0.007)</td>
</tr>
<tr>
<td>Chief complaint cluster: neurological</td>
<td>N/A</td>
<td>−0.038 (0.009)</td>
<td>−0.039 (0.009)</td>
</tr>
<tr>
<td>NLP cluster: eye, ENT, or skin</td>
<td>−0.134 (0.009)</td>
<td>N/A</td>
<td>−0.061 (0.009)</td>
</tr>
<tr>
<td>Chief complaint cluster: eye, ENT, or skin</td>
<td>N/A</td>
<td>−0.302 (0.013)</td>
<td>−0.279 (0.013)</td>
</tr>
<tr>
<td>NLP cluster: urological or gynecological</td>
<td>0.055 (0.008)</td>
<td>N/A</td>
<td>0.006 (0.008)</td>
</tr>
<tr>
<td>Chief complaint cluster: urological or gynecological</td>
<td>N/A</td>
<td>0.193 (0.012)</td>
<td>0.187 (0.013)</td>
</tr>
<tr>
<td>NLP cluster: trauma</td>
<td>−0.129 (0.005)</td>
<td>N/A</td>
<td>−0.098 (0.005)</td>
</tr>
<tr>
<td>Chief complaint cluster: trauma</td>
<td>N/A</td>
<td>−0.011 (0.007)</td>
<td>0.013 (0.007)</td>
</tr>
<tr>
<td>NLP cluster: psychiatric</td>
<td>0.063 (0.009)</td>
<td>N/A</td>
<td>0.080 (0.010)</td>
</tr>
<tr>
<td>Chief complaint cluster: psychiatric</td>
<td>N/A</td>
<td>0.086 (0.012)</td>
<td>0.051 (0.013)</td>
</tr>
</tbody>
</table>

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.

Observation: 52,222; R² = 0.409.
Observation: 52,222; R² = 0.448.
Observation: 52,222; R² = 0.457.
NLP: natural language processing.
P < .01.
N/A: not applicable.
P < .05.
ENT: ear, nose, and throat.
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").

<table>
<thead>
<tr>
<th>Variable and model</th>
<th>$F_1$-score on ones</th>
<th>AUC$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inpatient</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1: NLP$^b$ clusters</td>
<td>0.57</td>
<td>0.82</td>
</tr>
<tr>
<td>Model 2: LS$^c$ clusters</td>
<td>0.57</td>
<td>0.83</td>
</tr>
<tr>
<td>Model 3: NLP+LS clusters</td>
<td>0.59</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>Low ESI score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1: NLP clusters</td>
<td>0.86</td>
<td>0.92</td>
</tr>
<tr>
<td>Model 2: LS clusters</td>
<td>0.84</td>
<td>0.90</td>
</tr>
<tr>
<td>Model 3: NLP+LS clusters</td>
<td>0.87</td>
<td>0.92</td>
</tr>
</tbody>
</table>

$^a$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 be 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


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
Cheligeer Cheligeer\textsuperscript{1,2}, PhD; Guosong Wu\textsuperscript{1,3}, PhD; Seungwon Lee\textsuperscript{1,2}, PhD; Jie Pan\textsuperscript{1,3}, PhD; Danielle A Southern\textsuperscript{1}, MSc; Elliot A Martin\textsuperscript{1,2}, PhD; Natalie Sapiro\textsuperscript{1}, MSc, RN; Cathy A Eastwood\textsuperscript{1,3}, RN, PhD; Hude Quan\textsuperscript{1,3}, MD, PhD; Yuan Xu\textsuperscript{1,3,4,5}, MD, PhD

\textsuperscript{1}Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
\textsuperscript{2}Provincial Research Data Services, Alberta Health Services, Calgary, AB, Canada
\textsuperscript{3}Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
\textsuperscript{4}Department of Oncology, University of Calgary, Calgary, AB, Canada
\textsuperscript{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

\textbf{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.

\textbf{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.

\textbf{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.

\textbf{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.

\textbf{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.

\textit{(JMIR Med Inform 2024;12:e48995)} doi:10.2196/48995

\textbf{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 extract and analyze EMR data, potentially causing delays in recording and detecting AEs [15,16].

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 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 (κ<0.60), retraining occurred until high agreement (κ>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 MPRs notes), 2-day (fall day + 1 day after), 3-day (fall day + 2 days after), and full hospitalization MPRs. All supervised data sets were labeled to

https://medinform.jmir.org/2024/1/e48995 JMIR Med Inform 2024 | vol. 12 | e48995 | p.157 (page number not for citation purposes)
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.
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 F1-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 $\chi^2$ 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 F1-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>
<thead>
<tr>
<th>Description</th>
<th>Total (n=4323)</th>
<th>Confirmed fall (n=142)</th>
<th>No fall (n=4181)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, median (IQRs)</td>
<td>62.0 (48.5-75.5)</td>
<td>71.0 (59.6-82.4)</td>
<td>61.0 (47.5-74.5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sex (male), n (%)</td>
<td>2169 (50.2)</td>
<td>73 (51.4)</td>
<td>2096 (50.1)</td>
<td>.77</td>
</tr>
<tr>
<td>ICU visit, n (%)</td>
<td>163 (3.8)</td>
<td>19 (13.4)</td>
<td>144 (3.4)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Length of hospital stay (days), median (IQRs)</td>
<td>3.0 (0.5-5.5)</td>
<td>12.0 (1.5-22.5)</td>
<td>3.0 (0.5-5.5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Hospitals, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital “A”</td>
<td>3548 (82.1)</td>
<td>104 (73.2)</td>
<td>3444 (82.4)</td>
<td>.04</td>
</tr>
<tr>
<td>Hospital “B”</td>
<td>651 (15.1)</td>
<td>34 (23.9)</td>
<td>617 (14.8)</td>
<td></td>
</tr>
<tr>
<td>Hospital “C”</td>
<td>124 (2.8)</td>
<td>4 (2.9)</td>
<td>120 (2.8)</td>
<td></td>
</tr>
</tbody>
</table>

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 F1-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% F1-score. The high PPV model achieved 97.5% accuracy, 85.7% PPV, and 27.9% sensitivity. The high F1-model balanced 66.7% sensitivity and 60.5% PPV to optimize 64.4% F1-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 F1-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.

<table>
<thead>
<tr>
<th>Category and model name</th>
<th>Sensitivity (%), (95% CI)</th>
<th>Specificity (%), (95% CI)</th>
<th>PPV&lt;sup&gt;a&lt;/sup&gt; (%), (95% CI)</th>
<th>NPV&lt;sup&gt;b&lt;/sup&gt; (%), (95% CI)</th>
<th>Accuracy (%), (95% CI)</th>
<th>F&lt;sub&gt;1&lt;/sub&gt;-score&lt;sup&gt;c&lt;/sup&gt; (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BERT&lt;sup&gt;d&lt;/sup&gt;-based models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AHN-BERT&lt;sup&gt;e&lt;/sup&gt; (high sensitivity)</td>
<td>97.7 (87.7-99.9)</td>
<td>81.8 (79.6-83.9)</td>
<td>15.6 (14.0-17.3)</td>
<td>99.9 (99.3-100.0)</td>
<td>82.3 (80.1-84.4)</td>
<td>26.8</td>
</tr>
<tr>
<td>AHN-BERT (high PPV)</td>
<td>27.9 (15.3-43.7)</td>
<td>99.8 (99.4-100.0)</td>
<td>85.7 (57.2-98.2)</td>
<td>97.6 (96.6-98.4)</td>
<td>97.5 (96.5-98.2)</td>
<td>42.1</td>
</tr>
<tr>
<td>AHN-BERT (high F&lt;sub&gt;1&lt;/sub&gt;)</td>
<td>66.7 (49.8-80.9)</td>
<td>99.0 (98.2-99.5)</td>
<td>60.5 (44.4-75.0)</td>
<td>98.7 (97.8-99.2)</td>
<td>97.7 (96.7-98.4)</td>
<td>63.4</td>
</tr>
<tr>
<td>BERT-uncased</td>
<td>79.1 (64.0-9.0)</td>
<td>61.4 (58.6-64.1)</td>
<td>6.6 (5.6-7.7)</td>
<td>98.8 (97.9-99.4)</td>
<td>62.0 (59.3-64.6)</td>
<td>12.1</td>
</tr>
<tr>
<td>BioClinical BERT</td>
<td>74.4 (58.8-86.5)</td>
<td>69.8 (67.2-72.4)</td>
<td>7.8 (6.5-9.3)</td>
<td>98.8 (97.9-99.3)</td>
<td>70.0 (67.4-72.5)</td>
<td>14.1</td>
</tr>
<tr>
<td><strong>Classical machine learning classifier</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support vector machine</td>
<td>76.7 (61.4-88.2)</td>
<td>85.0 (82.9-86.9)</td>
<td>14.9 (12.5-17.8)</td>
<td>99.1 (98.4-99.5)</td>
<td>84.7 (82.6-86.6)</td>
<td>25.0</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>74.4 (58.8-86.5)</td>
<td>86.4 (84.3-88.2)</td>
<td>15.8 (13.0-19.0)</td>
<td>99.0 (98.3-99.4)</td>
<td>86.0 (83.9-87.8)</td>
<td>26.0</td>
</tr>
<tr>
<td>Decision tree</td>
<td>51.2 (35.5-66.7)</td>
<td>93.9 (92.4-95.1)</td>
<td>22.2 (14.5-31.7)</td>
<td>98.3 (97.3-98.9)</td>
<td>92.4 (90.9-93.8)</td>
<td>31.0</td>
</tr>
<tr>
<td><strong>Rule-based classifier</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICD 10&lt;sup&gt;f&lt;/sup&gt;</td>
<td>27.9 (15.3-43.7)</td>
<td>92.3 (90.7-93.8)</td>
<td>11.1 (6.9-17.3)</td>
<td>97.4 (96.9-97.8)</td>
<td>90.2 (88.5-91.8)</td>
<td>15.9</td>
</tr>
</tbody>
</table>

<sup>a</sup>PPV: positive predictive value. The proportion of true positive results among all positive results.

<sup>b</sup>NPV: negative predictive value. The proportion of true negative results among all negative results.

<sup>c</sup>F<sub>1</sub>-score: a measure of a model’s accuracy that considers both sensitivity and PPV.

<sup>d</sup>BERT: Bidirectional Encoder Representations from Transformers.

<sup>e</sup>AHN-BERT: Alberta hospital notes-specific BERT.

<sup>f</sup>ICD 10: International Classification of Diseases, 10th Revision.

Figure 2. Performance metrics at varying thresholds: PPV, sensitivity, and F1-score. 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 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.

References

https://medinform.jmir.org/2024/1/e48995 JMIR Med Inform 2024 | vol. 12 | e48995 | p.163 (page number not for citation purposes)


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

https://medinform.jmir.org/2024/1/e48995 JMIR Med Inform 2024 | vol. 12 | e48995 | p.165 (page number not for citation purposes)
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

©Cheligeer Cheligeer, Guosong Wu, Seungwon Lee, Jie Pan, Danielle A Southern, Elliot A Martin, Natalie Sapiro, Cathy A Eastwood, Hude Quan, Yuan Xu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 30.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
A Patient Similarity Network (CHDmap) to Predict Outcomes After Congenital Heart Surgery: Development and Validation Study

Haomin Li1,*, PhD; Mengying Zhou1,2,*, MSc; Yuhan Sun1,2,*, BSc; Jian Yang1,2, BSc; Xian Zeng1,2, PhD; Yunxiang Qiu3, MD; Yuanyuan Xia3, MD; Zhijie Zheng3, MD; Jin Yu4, MD; Yuqing Feng1, MSc; Zhuo Shi5, MD; Ting Huang5, MD; Linhua Tan3, MD; Ru Lin5, MD; Jianhua Li5, MD; Xiangming Fan3, MD; Jingjing Ye4, MD; Huilong Duan3, PhD; Shanshan Shi3,*, MD; Qiang Shu5,*, MD

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.
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].
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].

\[ d_{\text{final}} = \alpha d_{\text{echocardiography}} + \beta d_{\text{diagnosis}} + \gamma d_{\text{clinical}} + \delta d_{\text{surgery}} \]  

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).
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 (e.g., 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 $\chi^2$ 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, F1-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. F1-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 F1-score also considers both of them. AUC provides an aggregate measure of the performance across all possible classification thresholds. The higher the accuracy, recall, F1-score, and AUC, the better the model’s performance is at distinguishing between the positive and negative classes.

<table>
<thead>
<tr>
<th>Case</th>
<th>Accuracy</th>
<th>Recall</th>
<th>F1-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.78</td>
<td>0.82</td>
<td>0.80</td>
<td>0.85</td>
</tr>
<tr>
<td>2</td>
<td>0.80</td>
<td>0.85</td>
<td>0.82</td>
<td>0.87</td>
</tr>
<tr>
<td>3</td>
<td>0.82</td>
<td>0.87</td>
<td>0.84</td>
<td>0.89</td>
</tr>
<tr>
<td>4</td>
<td>0.84</td>
<td>0.89</td>
<td>0.86</td>
<td>0.90</td>
</tr>
<tr>
<td>5</td>
<td>0.86</td>
<td>0.90</td>
<td>0.88</td>
<td>0.91</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>Table 1. Characteristics of patients with CHD used to generate CHDmap and in the test data set.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
</tr>
<tr>
<td>Gender (male), n (%)</td>
</tr>
<tr>
<td>Age (mo), median (IQR)</td>
</tr>
<tr>
<td>Height (cm), median (IQR)</td>
</tr>
<tr>
<td>Weight (kg), median (IQR)</td>
</tr>
<tr>
<td>Preoperative oxygen saturation (%), median (IQR)</td>
</tr>
<tr>
<td>Surgery time (min), median (IQR)</td>
</tr>
<tr>
<td>Cardiopulmonary bypass time (min), median (IQR)</td>
</tr>
<tr>
<td>Aortic cross-clamping time (min), median (IQR)</td>
</tr>
<tr>
<td>Duration of hospital stay (d), median (IQR)</td>
</tr>
<tr>
<td>Duration of ICU stay (d), median (IQR)</td>
</tr>
<tr>
<td>Diagnostic label, n (%)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Mechanical ventilation time (%), n (%)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Complication, n (%)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

$^a$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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Prediction of postoperative complications</th>
<th>Prediction of mechanical ventilation duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Recall</td>
</tr>
<tr>
<td>KNN$^b$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Origin+un-grade</td>
<td>0.832</td>
<td>0.438</td>
</tr>
<tr>
<td>Origin+grade</td>
<td>0.836</td>
<td>0.417</td>
</tr>
<tr>
<td>z-score+un-grade</td>
<td>0.828</td>
<td>0.458</td>
</tr>
<tr>
<td>z-score+grade</td>
<td>0.848</td>
<td>0.458</td>
</tr>
<tr>
<td>Combination+un-grade</td>
<td>0.836</td>
<td>0.500</td>
</tr>
<tr>
<td>Combination+grade</td>
<td>0.859</td>
<td>0.458</td>
</tr>
<tr>
<td>KNN+LR$^c$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Origin+un-grade</td>
<td>0.813</td>
<td>0.604</td>
</tr>
<tr>
<td>Origin+grade</td>
<td>0.813</td>
<td>0.667</td>
</tr>
<tr>
<td>z-score+un-grade</td>
<td>0.809</td>
<td>0.604</td>
</tr>
<tr>
<td>z-score+grade</td>
<td>0.813</td>
<td>0.646</td>
</tr>
<tr>
<td>Combination+un-grade</td>
<td>0.805</td>
<td>0.583</td>
</tr>
<tr>
<td>Combination+grade</td>
<td>0.805</td>
<td>0.604</td>
</tr>
<tr>
<td>$k$-Random+LR</td>
<td>0.809</td>
<td>0.500</td>
</tr>
<tr>
<td>Clinicians$^e$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0.875</td>
<td>0.396</td>
</tr>
<tr>
<td>C2</td>
<td>0.758</td>
<td>0.646</td>
</tr>
<tr>
<td>C3</td>
<td>0.840</td>
<td>0.208</td>
</tr>
<tr>
<td>Clinician average</td>
<td>0.824</td>
<td>0.417</td>
</tr>
<tr>
<td>C1+CHDmap</td>
<td>0.883</td>
<td>0.426</td>
</tr>
<tr>
<td>C2+CHDmap</td>
<td>0.816</td>
<td>0.5625</td>
</tr>
<tr>
<td>C3+CHDmap</td>
<td>0.852</td>
<td>0.313</td>
</tr>
<tr>
<td>Clinician+CHDmap average</td>
<td>0.850</td>
<td>0.434</td>
</tr>
</tbody>
</table>

$^a$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. 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


41. Download. CHDMap. URL: http://chdmap.nbscn.org/Help#download [accessed 2024-01-04]


43. Li et al. 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:
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
PMID: 38297829

© Haomin Li, Mengying Zhou, Yuhan Sun, Jian Yang, Xian Zeng, Yunxiang Qiu, Yuanyuan Xia, Zhijie Zheng, Jin Yu, Yuqing Feng, Zhuo Shi, Ting Huang, Linhua Tan, Ru Lin, Jianhua Li, Xiangming Fan, Jingjing Ye, Huiulong Duan, Shanshan Shi, Qiang Shu. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.1.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Prediction of Antibiotic Resistance in Patients With a Urinary Tract Infection: Algorithm Development and Validation

Nevruz İlhanlı1,2*, MSc; Se Yoon Park1,3,4*, MD, PhD; Jaewoong Kim1,3, MSc; Jee An Ryu1, BA; Ahmet Yardımçı2, PhD; Dukyong Yoon1,4,5, MD, PhD

1Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Yongin, Republic of Korea
2Department of Biostatistics and Medical Informatics, Akdeniz University, Antalya, Turkey
3Department of Hospital Medicine, Yongin Severance Hospital, Yonsei University College of Medicine, Yongin, Republic of Korea
4Center for Digital Health, Yongin Severance Hospital, Yonsei University Health System, Yongin, Republic of Korea
5Institute 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, carbapenem, TMP-SMX, and fluoroquinolone, 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

https://medinform.jmir.org/2024/1/e51326
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.
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 AUROC on the test set across all iterations for each antibiotic was used to classify UTI pathogens as either sensitive or resistant to each antibiotic. 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.

<table>
<thead>
<tr>
<th></th>
<th>Cephalosporin</th>
<th>TZP&lt;sup&gt;a&lt;/sup&gt;</th>
<th>TMP-SMX&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Fluoroquinolone</th>
<th>Carbapenem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples, n</td>
<td>3535</td>
<td>737</td>
<td>708</td>
<td>1582</td>
<td>1365</td>
</tr>
<tr>
<td>Admissions, n</td>
<td>396</td>
<td>366</td>
<td>374</td>
<td>571</td>
<td>392</td>
</tr>
<tr>
<td>Patients, n</td>
<td>390</td>
<td>360</td>
<td>368</td>
<td>557</td>
<td>386</td>
</tr>
<tr>
<td>Resistance, n (%)</td>
<td>1492 (42.2)</td>
<td>169 (22.9)</td>
<td>281 (39.7)</td>
<td>1014 (64.1)</td>
<td>142 (10.4)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>71.5 (14.4)</td>
<td>71.4 (14.4)</td>
<td>71.4 (14.4)</td>
<td>71.9 (14.4)</td>
<td>71.7 (14.3)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>2597 (73.5)</td>
<td>523 (71)</td>
<td>507 (71.6)</td>
<td>1013 (64)</td>
<td>994 (72.8)</td>
</tr>
<tr>
<td>Most common bacteria</td>
<td>1650 (46.7)</td>
<td>312 (42.3)</td>
<td>331 (46.7)</td>
<td>349 (22)</td>
<td>624 (45.7)</td>
</tr>
<tr>
<td>(Escherichia coli), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second-most common</td>
<td>556 (15.7)</td>
<td>109 (14.8)</td>
<td>111 (15.7)</td>
<td>305 (19.3)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>220 (16.1)</td>
</tr>
<tr>
<td>bacteria (Klebsiella pneumoniae), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third-most common</td>
<td>168 (4.7)</td>
<td>69 (9.4)</td>
<td>21 (3)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>180 (11.4)&lt;sup&gt;e&lt;/sup&gt;</td>
<td>83 (6.1)</td>
</tr>
<tr>
<td>bacteria (Pseudomonas aeruginosa), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>TZP: piperacillin-tazobactam.  
<sup>b</sup>TMP-SMX: trimethoprim-sulfamethoxazole.  
<sup>c</sup>The isolated bacterial specimen is Enterococcus faecium.  
<sup>d</sup>The isolated bacterial specimen is Citrobacter freundii.  
<sup>e</sup>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), 0.665 (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.

<table>
<thead>
<tr>
<th></th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUROC&lt;sup&gt;a&lt;/sup&gt; (95% CI)</td>
<td>PRAUC&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Cephalosporin</td>
<td>0.777 (0.775-0.779)</td>
<td>0.725</td>
</tr>
<tr>
<td>TZP&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.864 (0.862-0.867)</td>
<td>0.688</td>
</tr>
<tr>
<td>Carbapenem</td>
<td>0.877 (0.874-0.880)</td>
<td>0.539</td>
</tr>
<tr>
<td>TMP-SMX&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.881 (0.879-0.882)</td>
<td>0.829</td>
</tr>
<tr>
<td>Fluoroquinolone</td>
<td>0.884 (0.884-0.885)</td>
<td>0.938</td>
</tr>
</tbody>
</table>

<sup>a</sup>AUROC: area under the receiver operating characteristic curve.  
<sup>b</sup>PRAUC: precision-recall area under the curve.  
<sup>c</sup>TZP: piperacillin-tazobactam.  
<sup>d</sup>TMP-SMX: trimethoprim-sulfamethoxazole.

Important Features

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.

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
NI, 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. NI performed the formal analysis of the collected data and wrote the paper. NI, 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.
References


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

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
The Implementation of an Electronic Medical Record in a German Hospital and the Change in Completeness of Documentation: Longitudinal Document Analysis

Florian Wurster¹, MSc; Marina Beckmann¹, DPhil; Natalia Cecon-Stabel¹, MSc; Kerstin Dittmer¹, MA; Till Jes Hansen¹, MSc; Julia Jaschke², MSc; Juliane Köberlein-Neu², Prof Dr; Mi-Ran Okumu¹, MA; Carsten Rusniok¹, MA; Holger Pfaff³, Prof Dr; Ute Karbach¹, PD, Dr

¹Chair 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
²Center 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; $P=0.01$; $d=-0.428$). 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].
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, and others.
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.

<table>
<thead>
<tr>
<th>ICD Code</th>
<th>Values, n/N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dorsalgia (ICD-M54)</td>
<td>213/3147 (6.77)</td>
</tr>
<tr>
<td>Other spondylopathies (ICD-M48)</td>
<td>133/3147 (4.23)</td>
</tr>
<tr>
<td>Fracture of forearm (ICD-S52)</td>
<td>131/3147 (4.16)</td>
</tr>
<tr>
<td>Complications of internal orthopedic prosthetic devices, implants and grafts (ICD-T84)</td>
<td>166/2912 (5.7)</td>
</tr>
<tr>
<td>Dorsalgia (ICD-M54)</td>
<td>148/2912 (5.08)</td>
</tr>
<tr>
<td>Fracture of shoulder and upper arm (ICD-S42)</td>
<td>121/2912 (4.16)</td>
</tr>
<tr>
<td>Complications of internal orthopedic prosthetic devices, implants and grafts (ICD-T84)</td>
<td>164/3091 (5.3)</td>
</tr>
<tr>
<td>Dorsalgia (ICD-M54)</td>
<td>163/3091 (5.27)</td>
</tr>
<tr>
<td>Fracture of forearm (ICD-S52)</td>
<td>159/3091 (5.14)</td>
</tr>
</tbody>
</table>

aICD: International Classification of Diseases.
bFrequency of coded diagnosis.
cTotal 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 = −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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type of record</th>
<th>Chi-square (df)</th>
<th>P value</th>
<th>RR² (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood pressure</td>
<td>Paper (n=44), n (%)</td>
<td>37 (84)</td>
<td>120 (88.2)</td>
<td>0.5 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>66 (48.5)</td>
<td>37.1 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Body temperature</td>
<td>Paper (n=44), n (%)</td>
<td>36 (81.8)</td>
<td>117 (86)</td>
<td>0.5 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>66 (48.5)</td>
<td>37.1 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Paper (n=44), n (%)</td>
<td>44 (100)</td>
<td>66 (48.5)</td>
<td>37.1 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>102 (75)</td>
<td>29.8 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diet</td>
<td>Paper (n=44), n (%)</td>
<td>13 (29.6)</td>
<td>102 (75)</td>
<td>5.8 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>92 (67.7)</td>
<td>5.8 (1)</td>
<td>.02</td>
</tr>
<tr>
<td>Excretions</td>
<td>Paper (n=44), n (%)</td>
<td>38 (86.4)</td>
<td>92 (67.7)</td>
<td>5.8 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>116 (85.3)</td>
<td>54.5 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Height</td>
<td>Paper (n=44), n (%)</td>
<td>12 (27.3)</td>
<td>116 (85.3)</td>
<td>54.5 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>106 (77.9)</td>
<td>7.0 (1)</td>
<td>.008</td>
</tr>
<tr>
<td>Pain</td>
<td>Paper (n=44), n (%)</td>
<td>42 (95.4)</td>
<td>106 (77.9)</td>
<td>7.0 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>120 (88.2)</td>
<td>1.2 (1)</td>
<td>.28</td>
</tr>
<tr>
<td>Pulse</td>
<td>Paper (n=44), n (%)</td>
<td>36 (81.8)</td>
<td>120 (88.2)</td>
<td>1.2 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>13 (9.6)</td>
<td>0.1 (1)</td>
<td>.73</td>
</tr>
<tr>
<td>Reanimation status</td>
<td>Paper (n=44), n (%)</td>
<td>5 (11.4)</td>
<td>13 (9.6)</td>
<td>0.1 (1)</td>
</tr>
<tr>
<td></td>
<td>Electronic (n=136), n (%)</td>
<td>117 (86)</td>
<td>56.5 (1)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

aRR: relative risk.
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 analyze 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 ($n_{\text{perc}}=44$ vs $n_{\text{electronic}}=136$). 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 preimplementation 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


Abbreviations

- **EMR**: electronic medical record
- **HIS**: hospital information system
- **ICD**: International Classification of Diseases
- **STROBE**: Strengthening the Reporting of Observational Studies in Epidemiology

©Florian Wurster, Marina Beckmann, Natàlia Cecon-Stabel, Kerstin Dittmer, Till Jes Hansen, Julia Jaschke, Juliane Köberlein-Neu, Mi-Ran Okumu, Carsten Rusniok, Holger Pfaß, Ute Karbach. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

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.
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 Zhan1*, MMed; Liping Ding1*, BSc; Hui Li1, MMed; Aonan Su1, 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.
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.
Table 1. Potential failure modes with their causes and effects.

<table>
<thead>
<tr>
<th>Process</th>
<th>Failure modes</th>
<th>Reasons</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmission in the hospital information system</td>
<td>• Basic information transmission error&lt;br&gt;• Inpatient information transmission error&lt;br&gt;• Expenses information transmission error</td>
<td>• Data interface errors&lt;br&gt;• The original data on the front page are erroneous&lt;br&gt;• The DRG is erroneous&lt;br&gt;• Affects patients’ medical reimbursement</td>
<td></td>
</tr>
<tr>
<td>Front page filled by physicians</td>
<td>• Incorrectly filled-in medical information&lt;br&gt;• Incorrectly filled in other information</td>
<td>• Do not understand the filling criteria&lt;br&gt;• Do not fill in carefully&lt;br&gt;• Incomplete quality control reminders</td>
<td>The original data on the front page are erroneous&lt;br&gt;The DRG is erroneous&lt;br&gt;Affects patients’ medical reimbursement</td>
</tr>
<tr>
<td>Front page quality control by physicians</td>
<td>• Inadequate quality control&lt;br&gt;• Inaccurate quality control</td>
<td>• No emphasis on quality control&lt;br&gt;• Unfamiliar with quality control rules&lt;br&gt;• Complexity of quality control rules&lt;br&gt;• Lack of information assistance</td>
<td>The original data on the front page are erroneous&lt;br&gt;The DRG is wrong&lt;br&gt;Affects patients’ medical reimbursement</td>
</tr>
<tr>
<td>Transmission in the physicians’ EMR system</td>
<td>• Inconsistency between the received data in the EMR system and original data</td>
<td>• Data interface errors&lt;br&gt;• Encoding conversion error</td>
<td>The original data on the front page are wrong&lt;br&gt;The DRG is wrong&lt;br&gt;Errorous data persist&lt;br&gt;The DRG is wrong</td>
</tr>
<tr>
<td>Coders’ proofreading and coding</td>
<td>• No data errors were found&lt;br&gt;• Wrong code for diagnosis, surgery, or operation&lt;br&gt;• Restoration error&lt;br&gt;• Diagnostic and surgical operation codes do not meet the requirements of patients’ insurance</td>
<td>• Formal quality control rules are too simple&lt;br&gt;Lack of internal quality control reminders&lt;br&gt;Insufficient professional capacity of coders&lt;br&gt;Few training opportunities for coders&lt;br&gt;Inadequate communication between coders and doctors&lt;br&gt;The criteria are different between the requirements of patients’ insurance and front page</td>
<td></td>
</tr>
<tr>
<td>Front page quality control by EMR management</td>
<td>• Inadequate quality control&lt;br&gt;• Late feedback for the results of quality control</td>
<td>• Using a sampling model to conduct quality control&lt;br&gt;Insufficient professional capacity of quality control staff&lt;br&gt;Complexity of quality control rules&lt;br&gt;Lack of information assistance</td>
<td>Unable to find all errors on the front page&lt;br&gt;Errorous data persist&lt;br&gt;The DRG is wrong</td>
</tr>
</tbody>
</table>

aDRG: diagnosis-related group.<br>bEMR: 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.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Severity</th>
<th>Criteria for risk severity</th>
<th>Frequency</th>
<th>Criteria for risk frequency</th>
<th>Detectability</th>
<th>Criteria for risk detectability</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Very high</td>
<td>Make the score of the front page of the EMR below 20</td>
<td>Extremely high</td>
<td>Every time</td>
<td>Very low</td>
<td>Cannot be detected</td>
</tr>
<tr>
<td>9</td>
<td>Very high</td>
<td>Make the score of the front page of the EMR between 20 and 30</td>
<td>Very high</td>
<td>Almost every time</td>
<td>Very low</td>
<td>Hard to detect</td>
</tr>
<tr>
<td>8</td>
<td>High</td>
<td>Make the score of the front page of the EMR between 30 and 40</td>
<td>Very high</td>
<td>One time every half day</td>
<td>Low</td>
<td>Seldom detected</td>
</tr>
<tr>
<td>7</td>
<td>High</td>
<td>Make the score of the front page of the EMR between 40 and 50</td>
<td>High</td>
<td>More than one time every day</td>
<td>Low</td>
<td>Seldom detected</td>
</tr>
<tr>
<td>6</td>
<td>Middle</td>
<td>Make the score of the front page of the EMR between 50 and 60</td>
<td>High</td>
<td>More than one time every week</td>
<td>Middle</td>
<td>Easy to be detected</td>
</tr>
<tr>
<td>5</td>
<td>Middle</td>
<td>Make the score of the front page of the EMR between 60 and 70</td>
<td>Middle</td>
<td>More than one time every month</td>
<td>Middle</td>
<td>Easy to be detected</td>
</tr>
<tr>
<td>4</td>
<td>Middle</td>
<td>Make the score of the front page of the EMR between 70 and 80</td>
<td>Middle</td>
<td>More than one time every year</td>
<td>High</td>
<td>Very easy to be detected</td>
</tr>
<tr>
<td>3</td>
<td>Low</td>
<td>Make the score of the front page of the EMR between 80 and 90</td>
<td>Low</td>
<td>One time every year</td>
<td>High</td>
<td>Very easy to be detected</td>
</tr>
<tr>
<td>2</td>
<td>Low</td>
<td>Make the score of the front page of the EMR between 90 and 100</td>
<td>Very low</td>
<td>Less than one time every year</td>
<td>High</td>
<td>Very easy to be detected</td>
</tr>
<tr>
<td>1</td>
<td>Very low</td>
<td>Does not affect the score of the front page of the EMR</td>
<td>Extremely low</td>
<td>Never</td>
<td>Very high</td>
<td>No failure modes</td>
</tr>
</tbody>
</table>

EMR: electronic medical record.

Table 3. Details of the expert panel.

<table>
<thead>
<tr>
<th>Position</th>
<th>Rating weight</th>
<th>Working experience (years)</th>
<th>Familiarity with FMEA</th>
<th>Rating weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician</td>
<td>High</td>
<td>&gt;20</td>
<td>General</td>
<td>9/10</td>
</tr>
<tr>
<td>Physician</td>
<td>Middle</td>
<td>10-20</td>
<td>General</td>
<td>7/10</td>
</tr>
<tr>
<td>Physician</td>
<td>Middle</td>
<td>1-5</td>
<td>Familiar</td>
<td>6/10</td>
</tr>
<tr>
<td>Physician</td>
<td>Primary</td>
<td>6-10</td>
<td>General</td>
<td>5/10</td>
</tr>
<tr>
<td>Coder</td>
<td>High</td>
<td>10-20</td>
<td>Familiar</td>
<td>9/10</td>
</tr>
<tr>
<td>Coder</td>
<td>Middle</td>
<td>6-10</td>
<td>Familiar</td>
<td>7/10</td>
</tr>
<tr>
<td>Quality control staff</td>
<td>High</td>
<td>10-20</td>
<td>Not very familiar</td>
<td>7/10</td>
</tr>
<tr>
<td>Administrator</td>
<td>High</td>
<td>&gt;20</td>
<td>Not very familiar</td>
<td>8/10</td>
</tr>
<tr>
<td>Administrator</td>
<td>Middle</td>
<td>10-20</td>
<td>Familiar</td>
<td>8/10</td>
</tr>
<tr>
<td>Information engineer</td>
<td>Primary</td>
<td>1-5</td>
<td>Familiar</td>
<td>5/10</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Why</th>
<th>What</th>
<th>How and When</th>
<th>Where</th>
<th>Who</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revamp the information</td>
<td>The rules for quality control are inadequate</td>
<td>Establish a set of intelligent front pages in the EMR management system</td>
<td>Added quality control rules from May to September 2021</td>
<td>EMR center and information center</td>
</tr>
<tr>
<td></td>
<td>Low accuracy in manual quality control</td>
<td></td>
<td>Embedded quality control systems into the physician's system and front page system from July to October 2021</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The criteria are different between the requirements of patients' insurance and front page details</td>
<td></td>
<td>Integrated medical insurance system and the front page of the EMR from July to December 2021</td>
<td></td>
</tr>
<tr>
<td>Revamp quality management</td>
<td>Formal quality control rules are simple</td>
<td>Improve the correctness of the front page by physician</td>
<td>From January to May 2021</td>
<td>EMR center and clinical departments</td>
</tr>
<tr>
<td></td>
<td>Insufficient professional capacity of coders</td>
<td>Improve the professional capacity of coders</td>
<td>Conducted special training seminars</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate communication between coders and physician</td>
<td>Establish effective communication channels</td>
<td>Prepared a quality and information management manual</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lack of a continuous improvement process</td>
<td>Form a continuous improvement process</td>
<td>Implemented professional training regularly</td>
<td></td>
</tr>
</tbody>
</table>

aEMR: 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 ($\chi^2=269.6; P<.001$); inpatient information ($\chi^2=175.9; P<.001$); diagnosis, surgery, and operation code ($\chi^2=32.9; P<.001$); and the overall accuracy rate of the front page ($\chi^2=239.3; P<.001$) and the integrity rate of the front page ($\chi^2=110.4; P<.001$) increased significantly (Table 6).
### Comparison of the risk analysis before and after failure mode and effects analysis model improvement.

<table>
<thead>
<tr>
<th>Process, failure modes</th>
<th>Before FMEA&lt;sup&gt;a&lt;/sup&gt;</th>
<th>After FMEA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Severity</td>
<td>Frequency</td>
</tr>
<tr>
<td><strong>Information import from HIS&lt;sup&gt;b&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic information transmission error</td>
<td>3.1</td>
<td>3.4</td>
</tr>
<tr>
<td>Inpatient information transmission error</td>
<td>3.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Expenses information transmission error</td>
<td>2.4</td>
<td>2.1</td>
</tr>
<tr>
<td><strong>Front page filled by physicians</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incorrectly filled-in medical information</td>
<td>6.1</td>
<td>6.6</td>
</tr>
<tr>
<td>Incorrectly filled-in other information</td>
<td>2.7</td>
<td>5.8</td>
</tr>
<tr>
<td><strong>Front page quality control by physicians</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inadequate quality control</td>
<td>6.5</td>
<td>6.4</td>
</tr>
<tr>
<td>Inaccurate quality control</td>
<td>6.1</td>
<td>6.5</td>
</tr>
<tr>
<td><strong>Information import from the physician’s EMR&lt;sup&gt;c&lt;/sup&gt; system</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inconsistency between the received data in EMR system and the original data</td>
<td>3.6</td>
<td>3.3</td>
</tr>
<tr>
<td><strong>Coders’ proofreading and coding</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No data errors were found</td>
<td>5.2</td>
<td>6.3</td>
</tr>
<tr>
<td>Wrong code for diagnosis, surgery, or operation</td>
<td>6.7</td>
<td>6.6</td>
</tr>
<tr>
<td>Restoration error</td>
<td>6.6</td>
<td>6.1</td>
</tr>
<tr>
<td><strong>Front page quality control by EMR management</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inadequate quality control</td>
<td>6.1</td>
<td>6.5</td>
</tr>
<tr>
<td>Late feedback of quality control results</td>
<td>6.1</td>
<td>6.4</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<sup>a</sup>FMEA: failure mode and effects analysis.

<sup>b</sup>HIS: hospital information system.

<sup>c</sup>EMR: electronic medical record.

<sup>d</sup>N/A: not applicable.
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...
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 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


Abbreviations
EMR: electronic medical record
FMEA: failure mode and effects analysis
RPN: risk priority number
SWOT: strengths, weaknesses, opportunities, and threats

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

©Siyi Zhan, Liping Ding, Hui Li, Aonan Su. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 19.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium,
provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Real-World Data Quality Framework for Oncology Time to Treatment Discontinuation Use Case: Implementation and Evaluation Study

Boshu Ru¹, PhD; Arthur Sillah¹, MPH, PhD; Kaushal Desai¹, MS, PhD; Sheenu Chandwani¹, MPH, PhD; Lixia Yao¹, PhD; Smita Kothari¹, 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 anticaner 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 (\(\text{date of last recorded dose} - \text{date of first recorded dose} + 1\) 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 $\geq 120$ 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).

<table>
<thead>
<tr>
<th>Operational steps to ascertain rwTTD and type of data category</th>
<th>Commonly used data elements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identify records of the drug of interest</strong></td>
<td></td>
</tr>
<tr>
<td>SACT drug</td>
<td>Drug_name, NDC(^a), HCPCS(^b) code, RxNorm code</td>
</tr>
<tr>
<td>SACT administration</td>
<td>Drug administration date(^c)</td>
</tr>
<tr>
<td>SACT order</td>
<td>Drug order date(^d)</td>
</tr>
<tr>
<td><strong>Identify discontinuation date from subsequent LOT start date</strong></td>
<td></td>
</tr>
<tr>
<td>LOT</td>
<td>LOT name(^e)</td>
</tr>
<tr>
<td>LOT</td>
<td>LOT number</td>
</tr>
<tr>
<td>LOT</td>
<td>LOT start date</td>
</tr>
<tr>
<td>LOT</td>
<td>LOT end date</td>
</tr>
<tr>
<td><strong>If no subsequent LOT, identify discontinuation date from patient death record during treatment</strong></td>
<td>Vital status or date of death</td>
</tr>
<tr>
<td>Mortality status</td>
<td></td>
</tr>
<tr>
<td><strong>If no date of death, identify discontinuation date by last follow-up date subheading</strong></td>
<td>Date of last follow-up(^f)</td>
</tr>
<tr>
<td>Last follow-up</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)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-administered 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 set B 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\(^a\).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Data set A (n=7366)</th>
<th>Data set B (n=11,386)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1723 (23.4)</td>
<td>3408 (29.9)</td>
</tr>
<tr>
<td>Male</td>
<td>5643 (76.6)</td>
<td>7967 (70)</td>
</tr>
<tr>
<td>Missing or unknown</td>
<td>0 (0)</td>
<td>11 (0.1)</td>
</tr>
<tr>
<td><strong>Age at first advanced HNC diagnosis (y), median (IQR)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;18</td>
<td>0 (0)</td>
<td>31 (0.27)</td>
</tr>
<tr>
<td>18-44</td>
<td>187 (2.53)</td>
<td>688 (6.04)</td>
</tr>
<tr>
<td>45-64</td>
<td>3402 (46.19)</td>
<td>5955 (52.3)</td>
</tr>
<tr>
<td>65-88</td>
<td>3777 (51.28)</td>
<td>4111 (36.11)</td>
</tr>
<tr>
<td>≥89</td>
<td>0 (0)</td>
<td>6 (0.05)</td>
</tr>
<tr>
<td>Missing or unknown</td>
<td>0 (0)</td>
<td>595 (5.23)</td>
</tr>
<tr>
<td><strong>Race or ethnicity, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>N/A(^b)</td>
<td>40 (0.35)</td>
</tr>
<tr>
<td>Asian</td>
<td>103 (1.4)</td>
<td>165 (1.45)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>487 (6.61)</td>
<td>1250 (10.98)</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>13 (0.18)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Native Hawaiian or other Pacific Islander</td>
<td>N/A(^b)</td>
<td>6 (0.05)</td>
</tr>
<tr>
<td>White</td>
<td>4939 (67.05)</td>
<td>9239 (81.14)</td>
</tr>
<tr>
<td>Missing</td>
<td>650 (8.82)</td>
<td>686 (6.02)</td>
</tr>
<tr>
<td>Other race</td>
<td>1174 (15.94)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>AJCC(^c) stage at first HNC diagnosis, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2 (0.03)</td>
<td>28 (0.25)</td>
</tr>
<tr>
<td>I</td>
<td>419 (5.69)</td>
<td>603 (5.3)</td>
</tr>
<tr>
<td>II</td>
<td>505 (6.86)</td>
<td>542 (4.76)</td>
</tr>
<tr>
<td>III</td>
<td>929 (12.61)</td>
<td>798 (7.01)</td>
</tr>
<tr>
<td>IV</td>
<td>4330 (58.78)</td>
<td>4978 (43.72)</td>
</tr>
<tr>
<td>Missing or unknown</td>
<td>1181 (16.03)</td>
<td>4437 (38.97)</td>
</tr>
<tr>
<td><strong>Year of first advanced HNC diagnosis, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Before 2006</td>
<td>0 (0)</td>
<td>1245 (10.9)</td>
</tr>
<tr>
<td>2006-2009</td>
<td>0 (0)</td>
<td>1537 (13.5)</td>
</tr>
<tr>
<td>2010-2012</td>
<td>1068 (14.5)</td>
<td>2721 (23.9)</td>
</tr>
<tr>
<td>2013-2018</td>
<td>5435 (73.8)</td>
<td>5577 (49.0)</td>
</tr>
<tr>
<td>2019 or later</td>
<td>863 (11.7)</td>
<td>306 (2.7)</td>
</tr>
</tbody>
</table>

\(^a\)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\(^c\).

<table>
<thead>
<tr>
<th>SACT data quality checks</th>
<th>Data set A</th>
<th>Data set B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1: patients with any SACT drug administration or order record after the advanced HNC diagnosis date, n (%)(^d)</td>
<td>5592 (75.9)</td>
<td>4411 (38.7)</td>
</tr>
<tr>
<td><strong>Task 2: SACT drug records with missing drug identity (name and code) information</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value, n (%)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Normalization of medication name</td>
<td>Normalized generic name</td>
<td>RxNorm ingredient level</td>
</tr>
<tr>
<td>Task 3: patients with target SACT administration date after the advanced HNC diagnosis date, 2015 onward, % (n/N)(^f)</td>
<td>24.96 (1200/4808)</td>
<td>5.92 (237/4003)</td>
</tr>
<tr>
<td>Task 4: SACT drug administration records with complete administration date, n (%)</td>
<td>425,505 (100)</td>
<td>37,662 (100)</td>
</tr>
<tr>
<td>Task 5: gap (in d) between the target SACT drug administration dates, median (IQR; range)</td>
<td>21 (21-21; 1-113)</td>
<td>21 (11-21; 1-824)</td>
</tr>
<tr>
<td>Task 6: patients with target SACT order date after the advanced HNC diagnosis date</td>
<td>N/A(^g)</td>
<td>N/A(^g)</td>
</tr>
<tr>
<td>Task 7 SACT drug order records with complete days supply and refill information, n (%)</td>
<td>1732 (53.4)</td>
<td>N/A(^h)</td>
</tr>
<tr>
<td>Task 8: SACT drug order records with complete order date, n (%)(^i)</td>
<td>3241 (100)</td>
<td>8380 (100)</td>
</tr>
<tr>
<td>Task 9: distribution of gaps (in d) between target SACT drug order dates, normalized by days supply, refill, and cancelation record</td>
<td>N/A(^g)</td>
<td>N/A(^g)</td>
</tr>
</tbody>
</table>

\(^a\)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.
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.

<table>
<thead>
<tr>
<th>Task</th>
<th>Data set A</th>
<th>Data set B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task 10:</strong> number of patients initiating the target SACT by year and quarter</td>
<td>Figure 1A</td>
<td>Figure 1B</td>
</tr>
<tr>
<td><strong>Task 11:</strong> completeness of LOT information, n (%)(^c)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients with complete line number</td>
<td>5594 (75.94)</td>
<td>4589 (40.3)</td>
</tr>
<tr>
<td>Patients with complete line name</td>
<td>5594 (75.94)</td>
<td>N/A(^d,e)</td>
</tr>
<tr>
<td>Patients with complete line start date</td>
<td>5594 (75.94)</td>
<td>4589 (40.3)</td>
</tr>
<tr>
<td>Patients with complete line end date</td>
<td>5594 (75.94)</td>
<td>N/A(^e)</td>
</tr>
<tr>
<td><strong>Task 12:</strong> patients for whom the first LOT number after the advanced HNC(^f) diagnosis date was not 1, n (%)(^f)</td>
<td>0 (0)</td>
<td>434 (3.81)</td>
</tr>
<tr>
<td><strong>Task 13:</strong> distribution of LOT number at target SACT initiation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients who received the target SACT, n</td>
<td>1200</td>
<td>237</td>
</tr>
<tr>
<td>Line 1, n (%)</td>
<td>481 (40.08)</td>
<td>65 (27.43)</td>
</tr>
<tr>
<td>Line 2, n (%)</td>
<td>486 (40.5)</td>
<td>92 (38.82)</td>
</tr>
<tr>
<td>Line 3, n (%)</td>
<td>161 (13.42)</td>
<td>54 (22.78)</td>
</tr>
<tr>
<td>Line 4, n (%)</td>
<td>46 (3.83)</td>
<td>16 (6.75)</td>
</tr>
<tr>
<td>Line 5, n (%)</td>
<td>13 (1.83)</td>
<td>10 (4.22)</td>
</tr>
<tr>
<td>Lines 6-10, n (%)</td>
<td>13 (1.83)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Task 14:</strong> use of target SACT in (1L)^(g) before approval date</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IL monotherapy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients who received the target SACT, % (n/N)</td>
<td>78.37 (377/481)</td>
<td>67.69 (44/65)</td>
</tr>
<tr>
<td>First administration date(^h) in database</td>
<td>July 15, 2015</td>
<td>November 10, 2014</td>
</tr>
<tr>
<td>Cutoff date for the earliest 5% receipt</td>
<td>August 29, 2016</td>
<td>February 4, 2016</td>
</tr>
<tr>
<td>Cutoff date for the earliest 10% receipt</td>
<td>November 3, 2016</td>
<td>September 23, 2016</td>
</tr>
<tr>
<td>Cutoff date for the earliest 25% receipt</td>
<td>June 28, 2017</td>
<td>April 27, 2017</td>
</tr>
<tr>
<td><strong>Approved IL combination</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients who received the target SACT in approved (1L)^(g), % (n/N)</td>
<td>7.69 (37/481)</td>
<td>6.15 (4/65)</td>
</tr>
<tr>
<td>First administration date in database</td>
<td>December 18, 2018</td>
<td>July 1, 2019</td>
</tr>
<tr>
<td>Cutoff date for the earliest 5% receipt</td>
<td>February 18, 2019</td>
<td>N/A(^i)</td>
</tr>
<tr>
<td>Cutoff date for the earliest 10% receipt</td>
<td>April 2, 2019</td>
<td>N/A(^i)</td>
</tr>
<tr>
<td>Cutoff date for the earliest 25% receipt</td>
<td>July 9, 2019</td>
<td>N/A(^i)</td>
</tr>
<tr>
<td><strong>Task 15:</strong> patients with death record, n (%)(^j)</td>
<td>4695 (63.74)</td>
<td>3531 (31)</td>
</tr>
<tr>
<td><strong>Task 16:</strong> patients with multiple death records on different dates, n (%)</td>
<td>0 (0)</td>
<td>N/A(^j)</td>
</tr>
<tr>
<td><strong>Task 17:</strong> patients with clinical records showing health care activity after death date (n=4695), n (%)</td>
<td>1497 (31.88)</td>
<td>N/A(^j)</td>
</tr>
<tr>
<td>≥1 d after date of death</td>
<td>1436 (30.59)</td>
<td>N/A(^j)</td>
</tr>
<tr>
<td>≥3 d after date of death</td>
<td>1290 (27.48)</td>
<td>N/A(^j)</td>
</tr>
<tr>
<td>≥7 d after date of death</td>
<td>1002 (21.34)</td>
<td>N/A(^j)</td>
</tr>
<tr>
<td>≥30 d after date of death</td>
<td>79 (1.68)</td>
<td>N/A(^j)</td>
</tr>
</tbody>
</table>

\(^a\) 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.
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\textsuperscript{a} diagnosis date (task 18): follow-up data for patients with advanced HNC diagnosis on or after January 1, 2011.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Data set A (n=7366)</th>
<th>Data set B (n=7754)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Value, n (%)</td>
<td>Value, n (%)</td>
</tr>
<tr>
<td></td>
<td>Unique patient-date pairs, n</td>
<td>Unique patient-date pairs, n</td>
</tr>
<tr>
<td></td>
<td>Pairs per patient, n</td>
<td>Pairs per patient, n</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>6567 (89.15)</td>
<td>6893 (88.9)</td>
</tr>
<tr>
<td></td>
<td>60,178</td>
<td>370,671</td>
</tr>
<tr>
<td>Drug records\textsuperscript{b}</td>
<td>5840 (79.28)</td>
<td>6802 (87.72)</td>
</tr>
<tr>
<td></td>
<td>113,948</td>
<td>269,225</td>
</tr>
<tr>
<td>Laboratory records</td>
<td>6860 (93.13)</td>
<td>6403 (82.58)</td>
</tr>
<tr>
<td></td>
<td>179,177</td>
<td>147,314</td>
</tr>
<tr>
<td>Facility visit</td>
<td>7269 (98.68)</td>
<td>6838 (88.19)</td>
</tr>
<tr>
<td></td>
<td>274,714</td>
<td>392,175</td>
</tr>
<tr>
<td>Vital sign measurements</td>
<td>7254 (98.48)</td>
<td>6123 (78.97)</td>
</tr>
<tr>
<td></td>
<td>233,623</td>
<td>217,797</td>
</tr>
<tr>
<td>Nondrug medical procedure</td>
<td>N/A\textsuperscript{c}</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>6740 (86.92)</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>390,556</td>
</tr>
<tr>
<td>Genomic test</td>
<td>N/A</td>
<td>118 (1.52)</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>208</td>
</tr>
<tr>
<td>Biomarker test</td>
<td>440 (5.97)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>469</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>1.1</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>ECOG PS\textsuperscript{d}</td>
<td>5416 (73.53)</td>
<td>7.7</td>
</tr>
<tr>
<td></td>
<td>100,607</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>17.7</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\textsuperscript{a}HNC: head and neck cancer.  
\textsuperscript{b}Any drug, not just systemic anticancer therapies.  
\textsuperscript{c}N/A: not applicable.  
\textsuperscript{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\textsuperscript{a} diagnosis on or after January 1, 2011.

<table>
<thead>
<tr>
<th>Task</th>
<th>Data set A (n=7366)</th>
<th>Data set B (n=7754)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Value, n</td>
<td>Value, median</td>
</tr>
<tr>
<td></td>
<td>(IQR; range)</td>
<td>(SD)</td>
</tr>
<tr>
<td>Task 19: frequency of visits during target SACT\textsuperscript{b,e}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1L\textsuperscript{d} combination therapy</td>
<td>101</td>
<td>0.13 (0.07)</td>
</tr>
<tr>
<td></td>
<td>0.07-0.16; 0.02-0.33</td>
<td>19</td>
</tr>
<tr>
<td>1L monotherapy</td>
<td>358</td>
<td>0.07 (0.05)</td>
</tr>
<tr>
<td></td>
<td>0.05-0.08; 0.01-0.50</td>
<td>0.07 (0.05)</td>
</tr>
<tr>
<td>2L\textsuperscript{e} monotherapy</td>
<td>634</td>
<td>0.08 (0.06)</td>
</tr>
<tr>
<td></td>
<td>0.06-0.10; 0.01-0.95</td>
<td>0.08 (0.06)</td>
</tr>
<tr>
<td>All other</td>
<td>104</td>
<td>0.14 (0.08)</td>
</tr>
<tr>
<td></td>
<td>0.12 (0.09-0.17; 0.02-0.48)</td>
<td>0.14 (0.08)</td>
</tr>
<tr>
<td>Task 20: for patients still alive, gap (in d) from the last target SACT administration and last visit\textsuperscript{f}</td>
<td>708</td>
<td>128 (199)</td>
</tr>
<tr>
<td></td>
<td>28 (6-187; 0-1118)</td>
<td>128 (199)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}HNC: head and neck cancer.  
\textsuperscript{b}SACT: systemic anticancer therapy.  
\textsuperscript{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.  
\textsuperscript{d}1L: first line of therapy after the advanced HNC diagnosis date.  
\textsuperscript{e}2L+: second-line or later therapy.  
\textsuperscript{f}Limited to patients who (1) were still alive \(\geq\) 180 days after last receipt of target SACT and (2) received last dose of target SACT \(\geq\) 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.

References


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]


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]


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

© Boshu Ru, Arthur Sillah, Kaushal Desai, Sheenu Chandwani, Lixia Yao, Smita Kothari. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 06.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited.
complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.
Knowledge Graph for Breast Cancer Prevention and Treatment: Literature-Based Data Analysis Study

Shuyan Jin¹, MPH; Haobin Liang², MSc; Wenxia Zhang¹, PhD; Huan Li¹, MM

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 included in this study were retrieved and analyzed for breast cancer prevention and treatment.
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 (HER2)–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.

<table>
<thead>
<tr>
<th>Group</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breast cancer subtypes and stages (n=1387)</strong></td>
<td></td>
</tr>
<tr>
<td>Malignant neoplasm of breast</td>
<td>587 (42.3)</td>
</tr>
<tr>
<td>Triple-negative breast neoplasms</td>
<td>56 (4)</td>
</tr>
<tr>
<td>HER2&lt;sup&gt;+&lt;/sup&gt;-positive carcinoma of breast</td>
<td>54 (3.9)</td>
</tr>
<tr>
<td>Carcinoma breast stage IV</td>
<td>48 (3.5)</td>
</tr>
<tr>
<td>Breast cancer metastatic</td>
<td>47 (3.4)</td>
</tr>
<tr>
<td>Early-stage breast carcinoma</td>
<td>42 (3)</td>
</tr>
<tr>
<td>Malignant neoplasms</td>
<td>31 (2.2)</td>
</tr>
<tr>
<td>Neoplasm</td>
<td>30 (2.2)</td>
</tr>
<tr>
<td>Metastatic triple-negative breast carcinoma</td>
<td>26 (1.9)</td>
</tr>
<tr>
<td>High-risk cancer</td>
<td>24 (1.7)</td>
</tr>
<tr>
<td>Neoplasm metastasis</td>
<td>21 (1.5)</td>
</tr>
<tr>
<td>Advanced cancer</td>
<td>19 (1.4)</td>
</tr>
<tr>
<td>Advanced breast carcinoma</td>
<td>19 (1.4)</td>
</tr>
<tr>
<td>HER2-negative breast cancer</td>
<td>18 (1.3)</td>
</tr>
<tr>
<td>Locally advanced malignant neoplasm</td>
<td>17 (1.2)</td>
</tr>
<tr>
<td>Advanced malignant neoplasm</td>
<td>15 (1.1)</td>
</tr>
<tr>
<td>Nonsmall cell lung carcinoma</td>
<td>15 (1.1)</td>
</tr>
<tr>
<td>Noninfiltrating intraductal carcinoma</td>
<td>14 (1)</td>
</tr>
<tr>
<td>Locally advanced breast cancer</td>
<td>13 (0.9)</td>
</tr>
<tr>
<td>Breast cancer stage III</td>
<td>11 (0.8)</td>
</tr>
<tr>
<td><strong>Treatment of breast cancer (n=1387)</strong></td>
<td></td>
</tr>
<tr>
<td>Pharmacotherapy</td>
<td>267 (19.3)</td>
</tr>
<tr>
<td>Neoadjuvant therapy</td>
<td>88 (6.3)</td>
</tr>
<tr>
<td>Hormone therapy</td>
<td>68 (4.9)</td>
</tr>
<tr>
<td>Chemotherapy (adjuvant)</td>
<td>54 (3.9)</td>
</tr>
<tr>
<td>Therapeutic procedure</td>
<td>53 (3.8)</td>
</tr>
<tr>
<td>Radiation therapy</td>
<td>48 (3.5)</td>
</tr>
<tr>
<td>Treatment protocols</td>
<td>43 (3.1)</td>
</tr>
<tr>
<td>Adjuvant therapy</td>
<td>36 (2.6)</td>
</tr>
<tr>
<td>Breast-conserving surgery</td>
<td>35 (2.5)</td>
</tr>
<tr>
<td>First-line treatment</td>
<td>31 (2.2)</td>
</tr>
<tr>
<td>Single-agent therapy</td>
<td>27 (1.9)</td>
</tr>
<tr>
<td>Mastectomy</td>
<td>27 (1.9)</td>
</tr>
<tr>
<td>Operative surgical procedures</td>
<td>20 (1.4)</td>
</tr>
<tr>
<td>Interventional procedure</td>
<td>16 (1.2)</td>
</tr>
<tr>
<td>Radiotherapy (adjuvant)</td>
<td>14 (1)</td>
</tr>
<tr>
<td>Excision of axillary lymph nodes group</td>
<td>13 (0.9)</td>
</tr>
<tr>
<td>Combined modality therapy</td>
<td>12 (0.9)</td>
</tr>
<tr>
<td>Excision</td>
<td>11 (0.8)</td>
</tr>
<tr>
<td>Targeted therapy</td>
<td>11 (0.8)</td>
</tr>
</tbody>
</table>
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.
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 preventative 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_v1211e52210_app1.docx ]
Multimedia Appendix 2
Different subtypes and stages of breast cancer.

Multimedia Appendix 3
Treatments of breast cancer.

Multimedia Appendix 4
Drugs for breast cancer.

References


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
Enhancing Health Equity by Predicting Missed Appointments in Health Care: Machine Learning Study

Yi Yang¹, MSc; Samaneh Madanian¹, PhD; David Parry², PhD
¹Auckland University of Technology, Auckland, New Zealand
²Murdoch 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 km² 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.
**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 1.** Research flow and procedure. AI: artificial intelligence; 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.
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 ($\chi^2$) and their $P$ values were calculated to investigate whether different levels of a variable contributed differently to DNS events.

The confidence level ($\alpha=.05$) 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 ($\chi^2$) 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]:

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

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.

<table>
<thead>
<tr>
<th>Categorical variables</th>
<th>Chi-square statistic</th>
<th>Chi-square P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dna_history_count</td>
<td>118,461</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>is_last_appt_dna</td>
<td>77,600</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Clinictypedesc</td>
<td>35,201</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>age_bins</td>
<td>34,810</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>primaryethnicityethbroadgroup3</td>
<td>17,098</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>leadtime_bins</td>
<td>11,048</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>maritalstatus_group</td>
<td>10,527</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>visit_type_group</td>
<td>3525</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>visittime_bin</td>
<td>3447</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>patcurrentdomiciledenitionindex</td>
<td>2655</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>is_multiple_appt_same_day</td>
<td>1913</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>op_prioritycode_group</td>
<td>1,496</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>is_working_day_ind</td>
<td>1,244</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Gender</td>
<td>4</td>
<td>.06</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Models and hyperparameters</th>
<th>R package</th>
<th>Range</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>penalty</td>
<td>glmnet</td>
<td>1e-10-1</td>
<td>Total amount of regularization used to prevent overfit and underfit</td>
</tr>
<tr>
<td>Random forest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trees</td>
<td>ranger</td>
<td>300-1000</td>
<td>Number of trees in the forest</td>
</tr>
<tr>
<td>Min_n</td>
<td>ranger</td>
<td>3-10</td>
<td>Minimum amount of data to further split a node</td>
</tr>
<tr>
<td>Mtry</td>
<td>ranger</td>
<td>3-5</td>
<td>Maximum number of features that will be randomly sampled to split a node</td>
</tr>
<tr>
<td>XGBoost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trees</td>
<td>xgboost</td>
<td>300-1000</td>
<td>Number of trees in the forest</td>
</tr>
<tr>
<td>Min_n</td>
<td>xgboost</td>
<td>3-10</td>
<td>Minimum amount of data to further split a node</td>
</tr>
<tr>
<td>mtry</td>
<td>xgboost</td>
<td>3-10</td>
<td>Maximum number of features that will be randomly sampled to split a node</td>
</tr>
<tr>
<td>tree_depth</td>
<td>xgboost</td>
<td>3-12</td>
<td>Maximum depth of the tree</td>
</tr>
</tbody>
</table>

\(^{a}\text{XGBoost: Extreme Gradient Boosting.}\)

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 \((\text{dna\_flag})\) 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
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\textsuperscript{a} costs in 2020 at the MDHB\textsuperscript{b} hospital\textsuperscript{c}.

<table>
<thead>
<tr>
<th>Clinics</th>
<th>NP\textsuperscript{d} DNS count</th>
<th>NP DNS price</th>
<th>Total NP DNS cost</th>
<th>FU\textsuperscript{e} DNS count</th>
<th>FU DNS price</th>
<th>Total FU cost</th>
<th>Total DNS cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthopedics</td>
<td>301</td>
<td>$346</td>
<td>$104,143</td>
<td>745</td>
<td>$263</td>
<td>$196,299</td>
<td>$300,442</td>
</tr>
<tr>
<td>Diabetes</td>
<td>90</td>
<td>$452</td>
<td>$40,658</td>
<td>576</td>
<td>$307</td>
<td>$176,643</td>
<td>$217,302</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>221</td>
<td>$239</td>
<td>$52,776</td>
<td>874</td>
<td>$174</td>
<td>$152,322</td>
<td>$205,099</td>
</tr>
<tr>
<td>Pediatric medicine</td>
<td>124</td>
<td>$600</td>
<td>$74,366</td>
<td>327</td>
<td>$395</td>
<td>$129,271</td>
<td>$203,637</td>
</tr>
<tr>
<td>Ear nose throat</td>
<td>253</td>
<td>$358</td>
<td>$90,571</td>
<td>367</td>
<td>$269</td>
<td>$98,744</td>
<td>$189,316</td>
</tr>
<tr>
<td>Gynecology</td>
<td>177</td>
<td>$403</td>
<td>$71,322</td>
<td>386</td>
<td>$280</td>
<td>$108,124</td>
<td>$179,446</td>
</tr>
<tr>
<td>Hematology</td>
<td>75</td>
<td>$632</td>
<td>$47,389</td>
<td>232</td>
<td>$348</td>
<td>$80,834</td>
<td>$128,223</td>
</tr>
<tr>
<td>Cardiology</td>
<td>109</td>
<td>$490</td>
<td>$53,397</td>
<td>245</td>
<td>$299</td>
<td>$73,259</td>
<td>$126,656</td>
</tr>
<tr>
<td>Radiation oncology</td>
<td>42</td>
<td>$505</td>
<td>$21,194</td>
<td>350</td>
<td>$293</td>
<td>$102,652</td>
<td>$123,846</td>
</tr>
<tr>
<td>General surgery</td>
<td>147</td>
<td>$387</td>
<td>$56,856</td>
<td>208</td>
<td>$309</td>
<td>$64,369</td>
<td>$121,225</td>
</tr>
<tr>
<td>Audiology</td>
<td>268</td>
<td>$214</td>
<td>$57,302</td>
<td>272</td>
<td>$214</td>
<td>$58,157</td>
<td>$115,459</td>
</tr>
<tr>
<td>Neurology</td>
<td>153</td>
<td>$617</td>
<td>$94,408</td>
<td>38</td>
<td>$400</td>
<td>$15,204</td>
<td>$109,612</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>68</td>
<td>$506</td>
<td>$34,393</td>
<td>186</td>
<td>$362</td>
<td>$67,401</td>
<td>$101,794</td>
</tr>
<tr>
<td>Medical oncology</td>
<td>18</td>
<td>$650</td>
<td>$11,703</td>
<td>229</td>
<td>$360</td>
<td>$82,327</td>
<td>$94,030</td>
</tr>
<tr>
<td>Dental</td>
<td>136</td>
<td>$244</td>
<td>$33,132</td>
<td>193</td>
<td>$244</td>
<td>$47,019</td>
<td>$80,151</td>
</tr>
<tr>
<td>Renal medicine</td>
<td>5</td>
<td>$559</td>
<td>$2,793</td>
<td>181</td>
<td>$344</td>
<td>$62,201</td>
<td>$64,995</td>
</tr>
<tr>
<td>Respiratory lab</td>
<td>38</td>
<td>$479</td>
<td>$18,192</td>
<td>121</td>
<td>$347</td>
<td>$42,021</td>
<td>$60,213</td>
</tr>
<tr>
<td>Obstetrics</td>
<td>101</td>
<td>$227</td>
<td>$22,906</td>
<td>143</td>
<td>$227</td>
<td>$32,431</td>
<td>$55,337</td>
</tr>
<tr>
<td>Respiratory sleep</td>
<td>20</td>
<td>$271</td>
<td>$5412</td>
<td>153</td>
<td>$271</td>
<td>$41,403</td>
<td>$46,815</td>
</tr>
<tr>
<td>Urology</td>
<td>65</td>
<td>$357</td>
<td>$23,178</td>
<td>85</td>
<td>$274</td>
<td>$23,253</td>
<td>$46,432</td>
</tr>
<tr>
<td>Dietetics</td>
<td>93</td>
<td>$175</td>
<td>$16,302</td>
<td>168</td>
<td>$175</td>
<td>$29,449</td>
<td>$45,751</td>
</tr>
<tr>
<td>General medicine</td>
<td>44</td>
<td>$517</td>
<td>$22,747</td>
<td>69</td>
<td>$322</td>
<td>$22,200</td>
<td>$44,948</td>
</tr>
<tr>
<td>Respiratory</td>
<td>39</td>
<td>$479</td>
<td>$18,671</td>
<td>70</td>
<td>$347</td>
<td>$24,309</td>
<td>$42,980</td>
</tr>
<tr>
<td>Dermatology</td>
<td>66</td>
<td>$316</td>
<td>$20,877</td>
<td>60</td>
<td>$236</td>
<td>$14,174</td>
<td>$35,051</td>
</tr>
<tr>
<td>Oral and maxillofacial</td>
<td>23</td>
<td>$296</td>
<td>$6799</td>
<td>124</td>
<td>$203</td>
<td>$25,185</td>
<td>$31,984</td>
</tr>
<tr>
<td>Endocrinology</td>
<td>25</td>
<td>$525</td>
<td>$13,127</td>
<td>34</td>
<td>$332</td>
<td>$11,284</td>
<td>$24,411</td>
</tr>
<tr>
<td>Rheumatology</td>
<td>18</td>
<td>$647</td>
<td>$11,643</td>
<td>31</td>
<td>$345</td>
<td>$10,693</td>
<td>$22,336</td>
</tr>
<tr>
<td>Plastic surgery (excluding burns)</td>
<td>18</td>
<td>$296</td>
<td>$5321</td>
<td>69</td>
<td>$203</td>
<td>$14,014</td>
<td>$19,335</td>
</tr>
<tr>
<td>GI\textsuperscript{f} endoscopy</td>
<td>0</td>
<td>$506</td>
<td>$0</td>
<td>52</td>
<td>$362</td>
<td>$18,843</td>
<td>$18,843</td>
</tr>
<tr>
<td>Community pediatrics</td>
<td>20</td>
<td>$600</td>
<td>$11,994</td>
<td>10</td>
<td>$395</td>
<td>$3953</td>
<td>$15,948</td>
</tr>
<tr>
<td>Infectious diseases</td>
<td>7</td>
<td>$738</td>
<td>$5169</td>
<td>19</td>
<td>$534</td>
<td>$10,152</td>
<td>$15,321</td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>1</td>
<td>$507</td>
<td>$507</td>
<td>29</td>
<td>$448</td>
<td>$12,990</td>
<td>$13,496</td>
</tr>
<tr>
<td>Podiatry</td>
<td>17</td>
<td>$207</td>
<td>$3522</td>
<td>47</td>
<td>$207</td>
<td>$9737</td>
<td>$13,259</td>
</tr>
<tr>
<td>Aged ATR\textsuperscript{g} health</td>
<td>18</td>
<td>$244</td>
<td>$4394</td>
<td>35</td>
<td>$244</td>
<td>$8545</td>
<td>$12,939</td>
</tr>
<tr>
<td>Under 65 ATR</td>
<td>3</td>
<td>$244</td>
<td>$732</td>
<td>5</td>
<td>$244</td>
<td>$1221</td>
<td>$1953</td>
</tr>
<tr>
<td>Cardiothoracic</td>
<td>0</td>
<td>$573</td>
<td>$0</td>
<td>4</td>
<td>$425</td>
<td>$1698</td>
<td>$1698</td>
</tr>
<tr>
<td>Anesthetics</td>
<td>9</td>
<td>0</td>
<td>$0</td>
<td>3</td>
<td>$0</td>
<td>$0</td>
<td>$0</td>
</tr>
</tbody>
</table>

\textsuperscript{a}DNS: Did Not Show.

\textsuperscript{b}MDHB: MidCentral District Health Board.

\textsuperscript{c}A currency exchange rate of NZD $1=US $0.61 is applicable for the listed costs.

https://medinform.jmir.org/2024/1/e48273
Table 4. Comparison of the ML\textsuperscript{a} models’ performance.

<table>
<thead>
<tr>
<th>Classifier and resampling strategy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC\textsuperscript{b}</th>
<th>Accuracy</th>
<th>Modeling cost</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Logistic regression</strong>&lt;br&gt;Undersampling (under_ratio=2)</td>
<td>0.5146</td>
<td>0.9227</td>
<td>0.8474</td>
<td>0.8897</td>
<td>Less than 1 hour (5 minutes)</td>
</tr>
<tr>
<td>Oversampling (over_ratio=0.5)</td>
<td>0.5091</td>
<td>0.9247</td>
<td>0.8592</td>
<td>0.8911</td>
<td>Less than 1 hour (14 minutes)</td>
</tr>
<tr>
<td><strong>Random forest</strong>&lt;br&gt;Undersampling (under_ratio=2)</td>
<td>0.8243</td>
<td>0.8524</td>
<td>0.9236</td>
<td>0.8501</td>
<td>Over 8 hours (8.4)</td>
</tr>
<tr>
<td>Oversampling (over_ratio=0.5)</td>
<td>0.5940</td>
<td>0.9260</td>
<td>0.9220</td>
<td>0.8990</td>
<td>Over 137 hours</td>
</tr>
<tr>
<td><strong>XGBoost\textsuperscript{c}</strong>&lt;br&gt;Undersampling (under_ratio=2)</td>
<td>0.8278</td>
<td>0.8490</td>
<td>0.9239</td>
<td>0.9117</td>
<td>Over 4 hours (4.8)</td>
</tr>
<tr>
<td>Oversampling (over_ratio=0.5)</td>
<td>0.8297</td>
<td>0.8549</td>
<td>0.9267</td>
<td>0.8529</td>
<td>Over 51 hours (51.83)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}ML: machine learning.  
\textsuperscript{b}AUC: area under the curve.  
\textsuperscript{c}XGBoost: Extreme Gradient Boosting.

**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.

<table>
<thead>
<tr>
<th>Algorithm and variable importance ranking</th>
<th>Undersampling model</th>
<th>Oversampling model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>dna_history_count</td>
<td>dna_history_count</td>
</tr>
<tr>
<td>1</td>
<td>is_working_day</td>
<td>is_working_day</td>
</tr>
<tr>
<td>2</td>
<td>is_last_appt_dna</td>
<td>is_multiple_appt_same_day</td>
</tr>
<tr>
<td>3</td>
<td>is_multiple_appt_same_day</td>
<td>is_last_appt_dna</td>
</tr>
<tr>
<td>4</td>
<td>lead_time</td>
<td>lead_time</td>
</tr>
<tr>
<td>Random forest</td>
<td>dna_history_count</td>
<td>dna_history_count</td>
</tr>
<tr>
<td>1</td>
<td>is_last_appt_dna</td>
<td>age_when_visit</td>
</tr>
<tr>
<td>2</td>
<td>lead_time</td>
<td>lead_time</td>
</tr>
<tr>
<td>3</td>
<td>age_bins</td>
<td>is_last_appt_dna</td>
</tr>
<tr>
<td>4</td>
<td>clinic_type_desc</td>
<td>clinic_type_desc</td>
</tr>
<tr>
<td>XGBoost(^a)</td>
<td>dna_history_count</td>
<td>dna_history_count</td>
</tr>
<tr>
<td>1</td>
<td>is_last_appt_dna</td>
<td>age_when_visit</td>
</tr>
<tr>
<td>2</td>
<td>age_when_visit</td>
<td>is_last_appt_dna</td>
</tr>
<tr>
<td>3</td>
<td>lead_time</td>
<td>ethnicity</td>
</tr>
<tr>
<td>4</td>
<td>lead_time</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Ethnicity</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)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_prioritycode 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_prioritycode 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 ethnic 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.

**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

<table>
<thead>
<tr>
<th>Deprivation level</th>
<th>Māori, n (%)</th>
<th>Pacific, n (%)</th>
<th>European, n (%)</th>
<th>Asian, n (%)</th>
<th>Other, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3113 (7)</td>
<td>293 (1)</td>
<td>37,314 (86)</td>
<td>2077 (5)</td>
<td>835 (2)</td>
</tr>
<tr>
<td>2</td>
<td>4951 (9)</td>
<td>429 (1)</td>
<td>46,405 (85)</td>
<td>1470 (3)</td>
<td>1071 (2)</td>
</tr>
<tr>
<td>3</td>
<td>6367 (13)</td>
<td>489 (1)</td>
<td>42,565 (84)</td>
<td>613 (1)</td>
<td>821 (2)</td>
</tr>
<tr>
<td>4</td>
<td>14,736 (14)</td>
<td>1747 (2)</td>
<td>84,728 (79)</td>
<td>4574 (4)</td>
<td>1593 (1)</td>
</tr>
<tr>
<td>5</td>
<td>14,400 (13)</td>
<td>3398 (3)</td>
<td>83,568 (77)</td>
<td>6015 (6)</td>
<td>1590 (1)</td>
</tr>
<tr>
<td>6</td>
<td>14,103 (15)</td>
<td>1759 (2)</td>
<td>74,351 (79)</td>
<td>2974 (3)</td>
<td>1248 (1)</td>
</tr>
<tr>
<td>7</td>
<td>13,442 (17)</td>
<td>3601 (5)</td>
<td>58,187 (75)</td>
<td>1858 (2)</td>
<td>870 (1)</td>
</tr>
<tr>
<td>8</td>
<td>36,843 (19)</td>
<td>5402 (3)</td>
<td>148,605 (75)</td>
<td>5434 (3)</td>
<td>1988 (1)</td>
</tr>
<tr>
<td>9</td>
<td>40,642 (24)</td>
<td>7324 (4)</td>
<td>111,319 (67)</td>
<td>5443 (3)</td>
<td>2442 (1)</td>
</tr>
<tr>
<td>10</td>
<td>31,998 (35)</td>
<td>6283 (7)</td>
<td>52,064 (56)</td>
<td>1610 (2)</td>
<td>521 (1)</td>
</tr>
</tbody>
</table>
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


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]


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]


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
©Yi Yang, Samaneh Madanian, David Parry. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 12.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/ , as well as this copyright and license information must be included.
Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study

Anne de Hond\textsuperscript{1,2,3}, MSc, PhD; Marieke van Buchem\textsuperscript{1,2,3}, MSc; Claudio Fanconi\textsuperscript{3,4}, MSc; Mohana Roy\textsuperscript{5}, MD; Douglas Blayney\textsuperscript{5}, MD; Ilse Kant\textsuperscript{1,6}, MSc, PhD; Ewout Steyerberg\textsuperscript{1,2}, MSc, PhD; Tina Hernandez-Boussard\textsuperscript{3,7,8}, MSc, PhD

\textsuperscript{1}Clinical AI Implementation and Research Lab, Leiden University Medical Centre, Leiden, Netherlands
\textsuperscript{2}Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands
\textsuperscript{3}Department of Medicine (Biomedical Informatics), Stanford Medicine, Stanford University, Stanford, CA, United States
\textsuperscript{4}Department of Electrical Engineering and Information Technology, ETH Zürich, Zürich, Switzerland
\textsuperscript{5}Department of Medical Oncology, Stanford Medicine, Stanford University, Stanford, CA, United States
\textsuperscript{6}Department of Digital Health, University Medical Centre Utrecht, Utrecht, Netherlands
\textsuperscript{7}Department of Biomedical Data Science, Stanford University, Stanford, CA, United States
\textsuperscript{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.

\textit{(JMIR Med Inform 2024;12:e51925)} doi: 10.2196/51925
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 chemotherapy 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 chemotherapy 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.

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
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.

Figure 1. Flowchart of cohort selection.
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 5x10-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 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.

<table>
<thead>
<tr>
<th></th>
<th>All (N=16,159)</th>
<th>No depression diagnosis within 1 month after onset of treatment (n=15,722, 97.3%)</th>
<th>Depression diagnosis within 1 month after onset of treatment (n=437, 2.7%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex (female), n (%)</td>
<td>8758 (53)</td>
<td>8296 (52.8)</td>
<td>272 (62.2)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>62 (15)</td>
<td>62 (15)</td>
<td>60 (14)</td>
</tr>
<tr>
<td><strong>Race and ethnicity, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>1870 (11.6)</td>
<td>1812 (11.5)</td>
<td>58 (13.3)</td>
</tr>
<tr>
<td>Non-Hispanic Asian</td>
<td>3582 (22.2)</td>
<td>3525 (22.4)</td>
<td>57 (13)</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>422 (2.6)</td>
<td>410 (2.6)</td>
<td>&lt;20 (&lt;5)</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>8864 (54.9)</td>
<td>8583 (54.6)</td>
<td>281 (64.3)</td>
</tr>
<tr>
<td>Other</td>
<td>1421 (8.8)</td>
<td>1392 (8.9)</td>
<td>29 (6.6)</td>
</tr>
<tr>
<td><strong>Insurance characteristics, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>8745 (54.1)</td>
<td>8496 (54)</td>
<td>249 (57)</td>
</tr>
<tr>
<td>Medicare</td>
<td>2590 (16)</td>
<td>2514 (16)</td>
<td>76 (17.4)</td>
</tr>
<tr>
<td>Medicaid</td>
<td>1917 (11.9)</td>
<td>1860 (11.8)</td>
<td>57 (13)</td>
</tr>
<tr>
<td>Other or not identified</td>
<td>2907 (18)</td>
<td>2852 (18.1)</td>
<td>55 (12.6)</td>
</tr>
<tr>
<td><strong>Treatment characteristics, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of hospitalizations one month prior to treatment</td>
<td>2083 (13)</td>
<td>2016 (13)</td>
<td>67 (15)</td>
</tr>
<tr>
<td>Number of emergency department visits 1 month prior to treatment</td>
<td>945 (6)</td>
<td>895 (6)</td>
<td>50 (11)</td>
</tr>
<tr>
<td>Charlson comorbidity score</td>
<td>6.9 (3.8)</td>
<td>6.9 (3.8)</td>
<td>6.9 (3.9)</td>
</tr>
<tr>
<td><strong>Tumor type, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>1772 (11)</td>
<td>1739 (11.1)</td>
<td>33 (7.6)</td>
</tr>
<tr>
<td>Lung</td>
<td>1001 (6.2)</td>
<td>973 (6.2)</td>
<td>28 (6.4)</td>
</tr>
<tr>
<td>Prostate</td>
<td>777 (4.8)</td>
<td>764 (4.9)</td>
<td>&lt;20 (&lt;5)</td>
</tr>
<tr>
<td>Colon and rectum</td>
<td>543 (3.4)</td>
<td>525 (3.3)</td>
<td>&lt;20 (&lt;5)</td>
</tr>
<tr>
<td>Non-Hodgkin lymphoma</td>
<td>535 (3.3)</td>
<td>527 (3.4)</td>
<td>&lt;20 (&lt;5)</td>
</tr>
<tr>
<td>Other</td>
<td>3459 (21.4)</td>
<td>3364 (21.4)</td>
<td>95 (21.7)</td>
</tr>
<tr>
<td>Missing</td>
<td>8072 (50)</td>
<td>7830 (49.8)</td>
<td>242 (55.4)</td>
</tr>
<tr>
<td><strong>Cancer stage, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>1492 (9.2)</td>
<td>1466 (9.3)</td>
<td>26 (5.9)</td>
</tr>
<tr>
<td>Stage II</td>
<td>1499 (9.3)</td>
<td>1468 (9.3)</td>
<td>31 (7.1)</td>
</tr>
<tr>
<td>Stage III</td>
<td>1329 (8.2)</td>
<td>1294 (8.2)</td>
<td>35 (8)</td>
</tr>
<tr>
<td>Stage IV</td>
<td>1758 (10.9)</td>
<td>1699 (10.8)</td>
<td>59 (13.5)</td>
</tr>
<tr>
<td>Missing</td>
<td>10081 (62.4)</td>
<td>9795 (62.3)</td>
<td>286 (65.4)</td>
</tr>
<tr>
<td><strong>Patient email information (1 month prior to treatment)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sent 1 or more emails, n (%)</td>
<td>4070 (25.2)</td>
<td>3943 (25.1)</td>
<td>127 (29.1)</td>
</tr>
<tr>
<td>Email length in words, mean (SD)</td>
<td>49 (35)</td>
<td>49 (35)</td>
<td>49 (35)</td>
</tr>
<tr>
<td>Sent emails at night, n (%)</td>
<td>308 (1.9)</td>
<td>296 (1.9)</td>
<td>&lt;20 (&lt;5)</td>
</tr>
<tr>
<td><strong>Mental health history, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of depression diagnosis</td>
<td>400 (2.5)</td>
<td>343 (2.2)</td>
<td>57 (13)</td>
</tr>
</tbody>
</table>
Depression diagnosis within 1 month after onset of treatment (n=437, 2.7%)  
No depression diagnosis within 1 month after onset of treatment (n=15,722, 97.3%)  

All (N=16,159)

Descriptive statistics

| History of antidepressant medication | 2219 (13.7) |
| History of mental health referral   | 2707 (16.8) |

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).

<table>
<thead>
<tr>
<th>Type of data</th>
<th>AUROC (95% CI)</th>
<th>Calibration intercept (95% CI)</th>
<th>Calibration slope (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured EHR data</td>
<td>0.74 (0.71 to 0.78)</td>
<td>0.07 (−0.09 to 0.24)</td>
<td>0.93 (0.77 to 1.09)</td>
</tr>
<tr>
<td>Patient emails</td>
<td>0.54 (0.52 to 0.56)</td>
<td>−0.02 (−0.18 to 0.14)</td>
<td>1.0 (0.52 to 1.48)</td>
</tr>
<tr>
<td>Structured EHR data and patient emails</td>
<td>0.74 (0.71 to 0.78)</td>
<td>0.07 (−0.09 to 0.24)</td>
<td>0.91 (0.76 to 1.07)</td>
</tr>
<tr>
<td>Clinician notes</td>
<td>0.5 (0.49 to 0.52)</td>
<td>−0.05 (−0.21 to 0.11)</td>
<td>0.94 (−1.32 to 3.2)</td>
</tr>
<tr>
<td>Structured EHR data and clinician notes</td>
<td>0.71 (0.68 to 0.75)</td>
<td>−0.09 (−0.25 to 0.07)</td>
<td>1.92 (1.57 to 2.28)</td>
</tr>
<tr>
<td>Structured EHR data, clinician notes, and patient emails</td>
<td>0.7 (0.67 to 0.73)</td>
<td>−0.16 (−0.32 to −0.0)</td>
<td>2.46 (1.98 to 2.93)</td>
</tr>
</tbody>
</table>

*AUROC: area under the receiver operating characteristics curve.*  
*EHR: electronic health record.*
### Table 3. Sensitivity, specificity, PPV\textsuperscript{a}, and NPV\textsuperscript{b} at different decision thresholds for predicting depression risk within 1 month after the onset of treatment (test data).

<table>
<thead>
<tr>
<th>Threshold and analysis</th>
<th>Structured EHR\textsuperscript{c} data</th>
<th>Patient emails</th>
<th>Structured EHR data and patient emails</th>
<th>Clinician notes</th>
<th>Structured EHR data and clinician notes</th>
<th>Structured EHR data, clinician notes, and patient emails</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>Sensitivity (n/N)</td>
<td>0.9 (140/156)</td>
<td>1.0 (156/156)</td>
<td>0.87 (136/156)</td>
<td>1.0 (156/156)</td>
<td>1.0 (156/156)</td>
</tr>
<tr>
<td></td>
<td>Specificity (n/N)</td>
<td>0.35 (1847/5231)</td>
<td>0.0 (0/5231)</td>
<td>0.37 (1915/5231)</td>
<td>0.0 (0/5231)</td>
<td>0.0 (0/5231)</td>
</tr>
<tr>
<td></td>
<td>PPV (n/N)</td>
<td>0.04 (140/3524)</td>
<td>0.03 (156/5387)</td>
<td>0.04 (136/3452)</td>
<td>0.03 (156/5387)</td>
<td>0.03 (156/5387)</td>
</tr>
<tr>
<td></td>
<td>NPV (n/N)</td>
<td>0.99 (1847/1863)</td>
<td>N/A\textsuperscript{d}</td>
<td>0.99 (1915/1935)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>3%</td>
<td>Sensitivity (n/N)</td>
<td>0.63 (98/156)</td>
<td>0.13 (20/156)</td>
<td>0.58 (90/156)</td>
<td>1.0 (156/156)</td>
<td>0.55 (86/156)</td>
</tr>
<tr>
<td></td>
<td>Specificity (n/N)</td>
<td>0.75 (3912/5231)</td>
<td>0.95 (4962/5231)</td>
<td>0.77 (4032/5231)</td>
<td>0.0 (0/5231)</td>
<td>0.82 (4293/5231)</td>
</tr>
<tr>
<td></td>
<td>PPV (n/N)</td>
<td>0.07 (98/1417)</td>
<td>0.07 (20/289)</td>
<td>0.07 (90/1289)</td>
<td>0.03 (156/5387)</td>
<td>0.08 (86/1024)</td>
</tr>
<tr>
<td></td>
<td>NPV (n/N)</td>
<td>0.99 (3912/3970)</td>
<td>0.97 (4962/5098)</td>
<td>0.98 (4032/4098)</td>
<td>N/A</td>
<td>0.98 (4293/4363)</td>
</tr>
<tr>
<td>10%</td>
<td>Sensitivity (n/N)</td>
<td>0.19 (29/156)</td>
<td>0.0 (0/156)</td>
<td>0.19 (30/156)</td>
<td>0.0 (0/156)</td>
<td>0.0 (0/156)</td>
</tr>
<tr>
<td></td>
<td>Specificity (n/N)</td>
<td>0.97 (5086/5231)</td>
<td>1.0 (5231/5231)</td>
<td>0.97 (5071/5231)</td>
<td>1.0 (5231/5231)</td>
<td>1.0 (5231/5231)</td>
</tr>
<tr>
<td></td>
<td>PPV (n/N)</td>
<td>0.17 (29/174)</td>
<td>N/A</td>
<td>0.16 (30/190)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>NPV (n/N)</td>
<td>0.98 (5086/5213)</td>
<td>0.97 (5231/5387)</td>
<td>0.98 (5071/5197)</td>
<td>0.97 (5231/5387)</td>
<td>0.97 (5231/5387)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}PPV: positive predictive value.

\textsuperscript{b}NPV: negative predictive value.

\textsuperscript{c}EHR: electronic health record.

\textsuperscript{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 (0.73-0.78; 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


31. Lamproidis, 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...


35. Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. ICML 2001;1:609-616 [FREE Full text]


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]


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
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 Li¹, MD; Sarah Mullin¹, PhD; Peter L Elkin¹, 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
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.

<table>
<thead>
<tr>
<th></th>
<th>Total (N=459), n (%)</th>
<th>Survived (n=422, 91.9%), n (%)</th>
<th>Expired (n=37, 8.1%), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gestational age (weeks)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>7 (1.5)</td>
<td>2 (28.6)</td>
<td>5 (71.4)</td>
</tr>
<tr>
<td>24</td>
<td>40 (8.7)</td>
<td>28 (70)</td>
<td>12 (30)</td>
</tr>
<tr>
<td>25</td>
<td>41 (8.9)</td>
<td>36 (87.8)</td>
<td>5 (12.2)</td>
</tr>
<tr>
<td>26</td>
<td>52 (11.3)</td>
<td>49 (94.2)</td>
<td>3 (5.8)</td>
</tr>
<tr>
<td>27</td>
<td>87 (19)</td>
<td>84 (96.6)</td>
<td>3 (3.4)</td>
</tr>
<tr>
<td>28</td>
<td>106 (23.1)</td>
<td>98 (92.5)</td>
<td>8 (7.5)</td>
</tr>
<tr>
<td>29</td>
<td>126 (27.5)</td>
<td>125 (99.2)</td>
<td>1 (0.8)</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>236 (51.4)</td>
<td>214 (90.7)</td>
<td>22 (9.3)</td>
</tr>
<tr>
<td>Female</td>
<td>223 (48.6)</td>
<td>208 (93.3)</td>
<td>15 (6.7)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>16 (3.5)</td>
<td>15 (93.7)</td>
<td>1 (6.3)</td>
</tr>
<tr>
<td>Black</td>
<td>69 (15)</td>
<td>62 (89.9)</td>
<td>7 (10.1)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>25 (5.4)</td>
<td>23 (92)</td>
<td>2 (8)</td>
</tr>
<tr>
<td>Native American</td>
<td>4 (0.9)</td>
<td>3 (75)</td>
<td>1 (25)</td>
</tr>
<tr>
<td>White</td>
<td>278 (60.1)</td>
<td>255 (91.7)</td>
<td>23 (8.3)</td>
</tr>
<tr>
<td>Other</td>
<td>25 (5.4)</td>
<td>23 (92)</td>
<td>2 (8)</td>
</tr>
<tr>
<td>Unknown</td>
<td>42 (9.2)</td>
<td>41 (97.6)</td>
<td>1 (2.4)</td>
</tr>
<tr>
<td><strong>Insurance</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>343 (74.7)</td>
<td>311 (90.7)</td>
<td>32 (9.3)</td>
</tr>
<tr>
<td>Government</td>
<td>116 (25.3)</td>
<td>113 (97.4)</td>
<td>3 (2.6)</td>
</tr>
<tr>
<td>Uninsured</td>
<td>2 (0.4)</td>
<td>0 (0)</td>
<td>2 (100)</td>
</tr>
<tr>
<td><strong>Family religion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catholic</td>
<td>100 (21.8)</td>
<td>91 (91)</td>
<td>9 (9)</td>
</tr>
<tr>
<td>Protestant</td>
<td>24 (5.2)</td>
<td>22 (91.7)</td>
<td>2 (8.3)</td>
</tr>
<tr>
<td>Jewish</td>
<td>16 (3.5)</td>
<td>15 (93.7)</td>
<td>1 (6.3)</td>
</tr>
<tr>
<td>Other</td>
<td>30 (6.5)</td>
<td>25 (83.3)</td>
<td>5 (16.7)</td>
</tr>
<tr>
<td>Unknown</td>
<td>289 (63)</td>
<td>269 (93.1)</td>
<td>20 (6.9)</td>
</tr>
<tr>
<td><strong>Type of delivery</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cesarean section</td>
<td>356 (77.6)</td>
<td>331 (93)</td>
<td>25 (7)</td>
</tr>
<tr>
<td>Vaginal delivery</td>
<td>103 (22.4)</td>
<td>91 (88.3)</td>
<td>12 (11.7)</td>
</tr>
<tr>
<td><strong>Pregnancy type</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singleton</td>
<td>247 (53.8)</td>
<td>230 (93.1)</td>
<td>17 (6.9)</td>
</tr>
<tr>
<td>Multiple</td>
<td>212 (46.2)</td>
<td>192 (90.6)</td>
<td>20 (9.4)</td>
</tr>
<tr>
<td><strong>Antenatal steroids</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Received</td>
<td>369 (80.4)</td>
<td>347 (94)</td>
<td>22 (6)</td>
</tr>
<tr>
<td>Partially received</td>
<td>71 (15.5)</td>
<td>65 (91.5)</td>
<td>6 (8.5)</td>
</tr>
<tr>
<td>Not received</td>
<td>19 (4.1)</td>
<td>14 (73.7)</td>
<td>5 (26.3)</td>
</tr>
</tbody>
</table>

Logistic regression, Naive 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\textsuperscript{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).

<table>
<thead>
<tr>
<th>Method</th>
<th>AUROC, mean (SD)</th>
<th>Precision, mean (SD)</th>
<th>F\textsubscript{1}-score</th>
<th>Log loss score</th>
<th>Brier score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.82 (0.16)</td>
<td>0.55 (0.25)</td>
<td>0.61</td>
<td>0.35</td>
<td>0.08</td>
</tr>
<tr>
<td>SVM\textsuperscript{b}</td>
<td>0.86 (0.13)</td>
<td>0.61 (0.24)</td>
<td>0.62</td>
<td>0.20</td>
<td>0.06</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.91 (0.07)</td>
<td>0.61 (0.22)</td>
<td>0.67</td>
<td>0.19</td>
<td>0.06</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.88 (0.10)</td>
<td>0.55 (0.25)</td>
<td>0.45</td>
<td>0.80</td>
<td>0.07</td>
</tr>
<tr>
<td>Neural network</td>
<td>0.84 (0.16)</td>
<td>0.65 (0.24)</td>
<td>0.67</td>
<td>0.30</td>
<td>0.05</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.74 (0.22)</td>
<td>0.39 (0.17)</td>
<td>0.40</td>
<td>3.90</td>
<td>0.25</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>0.64 (0.13)</td>
<td>0.24 (0.16)</td>
<td>0.34</td>
<td>1.74</td>
<td>0.07</td>
</tr>
</tbody>
</table>

\textsuperscript{a}AUROC: area under the receiver operating characteristic.

\textsuperscript{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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Chi-square (df)</th>
</tr>
</thead>
<tbody>
<tr>
<td>un_ins</td>
<td>Uninsured</td>
<td>22.8 (1)</td>
</tr>
<tr>
<td>breathing1</td>
<td>Initial breathing assessment</td>
<td>21.4 (2)</td>
</tr>
<tr>
<td>anomaly</td>
<td>Serious fetal anomaly</td>
<td>20.8 (1)</td>
</tr>
<tr>
<td>airway1</td>
<td>Initial type of airway or ventilation</td>
<td>17.1 (4)</td>
</tr>
<tr>
<td>religion_jehovahs</td>
<td>Religion Jehovah’s Witness</td>
<td>11.4 (1)</td>
</tr>
<tr>
<td>twintwin</td>
<td>Twin-twin transfusion syndrome</td>
<td>9.4 (1)</td>
</tr>
<tr>
<td>uncertain</td>
<td>Uncertain pregnancy dating</td>
<td>7.9 (1)</td>
</tr>
<tr>
<td>religion_other</td>
<td>Religion other</td>
<td>4.9 (1)</td>
</tr>
<tr>
<td>gov_ins</td>
<td>Medicaid or Medicare insurance</td>
<td>4.7 (1)</td>
</tr>
<tr>
<td>muscle1</td>
<td>Muscle tone</td>
<td>4.6 (4)</td>
</tr>
</tbody>
</table>
Table 4. Pearson correlation: correlation of continuous features with mortality.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FiO2_1</td>
<td>Initial amount of oxygen ventilation</td>
<td>0.28</td>
</tr>
<tr>
<td>BD1</td>
<td>Initial arterial blood gas base deficit</td>
<td>0.23</td>
</tr>
<tr>
<td>dopa1</td>
<td>Initial IV(^a) dopamine rate</td>
<td>0.20</td>
</tr>
<tr>
<td>temp1</td>
<td>Initial temperature</td>
<td>0.14</td>
</tr>
<tr>
<td>pCO2_1</td>
<td>Initial arterial blood gas pCO(_2)^b</td>
<td>0.13</td>
</tr>
<tr>
<td>G</td>
<td>Maternal gravidity</td>
<td>0.07</td>
</tr>
<tr>
<td>P</td>
<td>Maternal parity</td>
<td>0.06</td>
</tr>
<tr>
<td>PRBC1</td>
<td>Initial IV blood transfusion amount</td>
<td>0.06</td>
</tr>
<tr>
<td>maternal_age</td>
<td>Maternal age</td>
<td>0.05</td>
</tr>
<tr>
<td>gluc1</td>
<td>Initial glucose</td>
<td>0.03</td>
</tr>
<tr>
<td>bands1</td>
<td>Initial bands</td>
<td>0.03</td>
</tr>
<tr>
<td>multiple</td>
<td>Number of fetuses at delivery</td>
<td>0.02</td>
</tr>
<tr>
<td>pO2_1</td>
<td>Initial arterial blood gas pO(_2)^c</td>
<td>–0.01(^d)</td>
</tr>
<tr>
<td>wbc1</td>
<td>Initial white blood cells</td>
<td>–0.02</td>
</tr>
<tr>
<td>BPmean1</td>
<td>Initial mean blood pressure</td>
<td>–0.04</td>
</tr>
<tr>
<td>monos1</td>
<td>Initial monocytes</td>
<td>–0.05</td>
</tr>
<tr>
<td>HR1</td>
<td>Initial heart rate</td>
<td>–0.05</td>
</tr>
<tr>
<td>hct1</td>
<td>Initial hematocrit</td>
<td>–0.07</td>
</tr>
<tr>
<td>neuts1</td>
<td>Initial neutrophil count</td>
<td>–0.07</td>
</tr>
<tr>
<td>SaO2_1</td>
<td>Initial oxygen saturation</td>
<td>–0.20</td>
</tr>
<tr>
<td>birth_wt</td>
<td>Birth weight</td>
<td>–0.22</td>
</tr>
<tr>
<td>GA</td>
<td>Gestational age at birth</td>
<td>–0.32</td>
</tr>
<tr>
<td>apgar1</td>
<td>One-minute APGAR(^e) score</td>
<td>–0.32</td>
</tr>
<tr>
<td>apgar5</td>
<td>Five-minute APGAR score</td>
<td>–0.35</td>
</tr>
</tbody>
</table>

\(^a\)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.

<table>
<thead>
<tr>
<th>Logistic regression: positively associated with mortality</th>
<th>Logistic regression: negatively associated with mortality</th>
<th>Random forest</th>
<th>AdaBoost</th>
<th>SVMb</th>
<th>Neural network</th>
</tr>
</thead>
<tbody>
<tr>
<td>race_hispanic</td>
<td>GA</td>
<td>GA</td>
<td>neuts</td>
<td>activity1</td>
<td>dopa1</td>
</tr>
<tr>
<td>color1</td>
<td>race_unk</td>
<td>birth_wt</td>
<td>hct1</td>
<td>GA</td>
<td>activity1</td>
</tr>
<tr>
<td>anomaly</td>
<td>apgar1</td>
<td>SaO2_1</td>
<td>SaO2_1</td>
<td>HTN</td>
<td>multiple</td>
</tr>
<tr>
<td>race_asian</td>
<td>gov_ins</td>
<td>BD1</td>
<td>wbc</td>
<td>anomaly</td>
<td>uncertain</td>
</tr>
<tr>
<td>un_ins</td>
<td>activity1</td>
<td>apgar1</td>
<td>apgar1</td>
<td>breathL1</td>
<td>race_unk</td>
</tr>
<tr>
<td>dopa1</td>
<td>monos1</td>
<td>gluc1</td>
<td>monos1</td>
<td>breathR1</td>
<td>twintwin</td>
</tr>
<tr>
<td>abdomen1</td>
<td>breathL1</td>
<td>dopa1</td>
<td>temp1</td>
<td>twintwin</td>
<td>abnormal</td>
</tr>
<tr>
<td>pvt_ins</td>
<td>PRBC1</td>
<td>apgar5</td>
<td>HR1</td>
<td>birth_wt</td>
<td>muscle1</td>
</tr>
<tr>
<td>multiple</td>
<td>infert</td>
<td>FiO2_1</td>
<td>FiO2_1</td>
<td>antfont1</td>
<td>wbc1</td>
</tr>
<tr>
<td>FiO2_1</td>
<td>dm</td>
<td>neuts</td>
<td>bands1</td>
<td>cprefill1</td>
<td>abdomen1</td>
</tr>
</tbody>
</table>

The descriptions of variable names are present in Multimedia Appendix 1.

SVM: support vector machine.

Table 6. AUROCa, 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).

<table>
<thead>
<tr>
<th>Method</th>
<th>AUROC, mean (SD)</th>
<th>Precision, mean (SD)</th>
<th>F1-score</th>
<th>Log loss score</th>
<th>Brier score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.77 (0.14)</td>
<td>0.41 (0.18)</td>
<td>0.51</td>
<td>0.29</td>
<td>0.07</td>
</tr>
<tr>
<td>SVMb</td>
<td>0.76 (0.10)</td>
<td>0.37 (0.15)</td>
<td>0.46</td>
<td>0.25</td>
<td>0.07</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.80 (0.14)</td>
<td>0.54 (0.27)</td>
<td>0.59</td>
<td>0.22</td>
<td>0.06</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.75 (0.17)</td>
<td>0.44 (0.29)</td>
<td>0.54</td>
<td>0.27</td>
<td>0.07</td>
</tr>
<tr>
<td>Neural network</td>
<td>0.76 (0.11)</td>
<td>0.44 (0.18)</td>
<td>0.53</td>
<td>0.31</td>
<td>0.07</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.68 (0.21)</td>
<td>0.30 (0.11)</td>
<td>0.19</td>
<td>6.09</td>
<td>0.59</td>
</tr>
<tr>
<td>K-nearest neighbor</td>
<td>0.62 (0.12)</td>
<td>0.20 (0.12)</td>
<td>0.30</td>
<td>1.77</td>
<td>0.09</td>
</tr>
</tbody>
</table>

AUROC: area under the receiver operating characteristic.
SVM: support vector machine.

Table 7. Prepartum features of highest importance in various models, listed in order of importance.

<table>
<thead>
<tr>
<th>Logistic regression: positively associated with mortality</th>
<th>Logistic regression: negatively associated with mortality</th>
<th>Random forest</th>
<th>AdaBoost</th>
<th>SVMb</th>
<th>Neural network</th>
</tr>
</thead>
<tbody>
<tr>
<td>maternal_age</td>
<td>GA</td>
<td>GA</td>
<td>birth_wt</td>
<td>maternal_age</td>
<td>GA</td>
</tr>
<tr>
<td>anomaly</td>
<td>race_unk</td>
<td>birth_wt</td>
<td>maternal_age</td>
<td>steroids</td>
<td>steroids</td>
</tr>
<tr>
<td>un_ins</td>
<td>dm</td>
<td>maternal_age</td>
<td>GA</td>
<td>P</td>
<td>HTN</td>
</tr>
<tr>
<td>asthma</td>
<td>depression</td>
<td>anomaly</td>
<td>G</td>
<td>inferf</td>
<td>GA</td>
</tr>
<tr>
<td>religion_jehovahs</td>
<td>PTL</td>
<td>G</td>
<td>multiple</td>
<td>G</td>
<td>anomaly</td>
</tr>
<tr>
<td>pvt_ins</td>
<td>steroids</td>
<td>P</td>
<td>religion_unk</td>
<td>uncertain</td>
<td>twintwin</td>
</tr>
<tr>
<td>race_hispanic</td>
<td>gov_ins</td>
<td>un_ins</td>
<td>religion_unk</td>
<td>uncertain</td>
<td>twintwin</td>
</tr>
<tr>
<td>twintwin</td>
<td>HTN</td>
<td>steroids</td>
<td>sex</td>
<td>sex</td>
<td>sex</td>
</tr>
<tr>
<td>uncertain</td>
<td>infert</td>
<td>uncertain</td>
<td>anomaly</td>
<td>multiple</td>
<td>P</td>
</tr>
<tr>
<td>multiple</td>
<td>P</td>
<td>twintwin</td>
<td>P</td>
<td>anomaly</td>
<td>SVD</td>
</tr>
</tbody>
</table>

The descriptions of variable names are present in Multimedia Appendix 1.
SVM: support vector machine.

https://medinform.jmir.org/2024/1/e42271
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


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

https://medinform.jmir.org/2024/1/e42271
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
PMID: 38273795
Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review

Yuan Peng\textsuperscript{1}, MSc; Franziska Bathelt\textsuperscript{2}, Dr Rer Nat; Richard Gebler\textsuperscript{1}, MSc; Robert Gött\textsuperscript{1}, Dipl.-Ing.; Andreas Heidenreich\textsuperscript{4}, Dipl.-Biol., Dr Rer Nat; Elisa Henke\textsuperscript{1}, MSc; Dennis Kadioglu\textsuperscript{4,5}, MSc; Stephan Lorenz\textsuperscript{1}, MSc; Abishaa Vengadeswaran\textsuperscript{5}, MSc; Martin Sedlmayr\textsuperscript{1}, Dr Rer Nat, Prof Dr

\textsuperscript{1}Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany
\textsuperscript{2}Thiem-Research GmbH, Cottbus, Germany
\textsuperscript{3}Core Unit Datenintegrationszentrum, University Medicine Greifswald, Greifswald, Germany
\textsuperscript{4}Department for Information and Communication Technology (DICT), Data Integration Center (DIC), Goethe University Frankfurt, University Hospital, Frankfurt am Main, Germany
\textsuperscript{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.
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.
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).
Table 1. MDD\textsuperscript{a} approaches that are categorized as “other.”

<table>
<thead>
<tr>
<th>MDD approach type and publication</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UML\textsuperscript{b}-based</strong></td>
<td>Dhaouadi et al [46] UML class diagram is used for modeling the transformation process</td>
</tr>
<tr>
<td><strong>Graphic-based</strong></td>
<td>Dhaouadi et al [46] BPMN\textsuperscript{c} standard is used for modeling an ETL\textsuperscript{d} process</td>
</tr>
<tr>
<td><strong>Ad hoc formalisms-based</strong></td>
<td>Dhaouadi et al [46] Entity Mapping Diagram is used for representing ETL tasks</td>
</tr>
<tr>
<td><strong>MDA\textsuperscript{e}-based</strong></td>
<td>Dhaouadi et al [46] MDA is a multilayered framework with multiple submodules for separation of the specification of a functionality from its implementation</td>
</tr>
<tr>
<td><strong>Message-based</strong></td>
<td>Novak et al [51] “Normal message” contains information of mapping and transformation; “command message” configures the (execution) system</td>
</tr>
<tr>
<td><strong>Template-based</strong></td>
<td>McCarthy et al [58] A transformation template for each data source that manages the complex transformation process</td>
</tr>
<tr>
<td></td>
<td>Binding et al [62] A template contains the mapping patterns which is then used for querying in database</td>
</tr>
<tr>
<td><strong>Metadata-based\textsuperscript{f}</strong></td>
<td>Ozyurt and Grethe [36] Implementing a generic data transformation language to transform heterogeneous data from multiple sources to a common format</td>
</tr>
<tr>
<td></td>
<td>Tomingas et al [44] Metadata of the source and target stored in a knowledge and metadata repository</td>
</tr>
<tr>
<td></td>
<td>Suleykin and Panfilov [60] Metadata of the mapping path stored in a metadata management framework</td>
</tr>
</tbody>
</table>

\textsuperscript{a}MDD: metadata-driven.  
\textsuperscript{b}UML: unified modeling language.  
\textsuperscript{c}BPMN: Business Process Model Notation.  
\textsuperscript{d}ETL: Extract-Transform-Load.  
\textsuperscript{e}MDA: Model Driven Architecture.  
\textsuperscript{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 Description | MDD approaches | | |
|----------------------|----------------|----------------|
|                      | 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.
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


12. FHIR v4.0.1. HL7 International. URL: https://www.hl7.org/fhir/ [accessed 2022-04-05]


18. Leitfaden Basis DE (R4). HL7 FHIR Implementierungsleitfäden. URL: https://ig.fhir.de/basisprofile-de/stable/Home.html [accessed 2023-08-01]


20. US Core implementation guide. HL7 International. URL: https://www.hl7.org/fhir/us/core/ [accessed 2022-12-16]


27. Leitfaden Basis DE (R4). HL7 FHIR Implementierungsleitfäden. URL: https://ig.fhir.de/basisprofile-de/stable/Home.html [accessed 2023-08-01]


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